A model of intentional communication: AIRBUS (Asymmetric Intention Recognition with Bayesian Updating of Signals)

J. P. de Ruiter and Chris Cummins
Universität Bielefeld

Abstract

The rapid and fluent nature of human communicative interactions strongly suggests the existence of an online mechanism for intention recognition. In this paper, we discuss the requirements of such a system, drawing upon philosophical and empirical considerations, and propose a new mathematical model that addresses these requirements. Our model provides a way of integrating knowledge about both directions of the relationship between linguistic expressions and communicative intentions, through a rapid process of Bayesian update. It enables us to frame predictions about the process of intention recognition. Moreover, by equipping this model with a measure of entropy, we can also draw predictions about the circumstances under which utterance planning can start before the preceding utterance is completed, and about when explicit other-repair mechanisms must be invoked to resolve misalignment. We sketch an example of this model in operation and discuss its implications for a broader theory of communication.

Introduction

The ability to communicate effectively and flexibly with other humans is one of our species’ most impressive cognitive capacities. However, there are very few comprehensive theories that aim to address this capacity, and those that do are often sketchy and fail to capture the essential and unique aspects of human communication.

Most notably among these, Shannon’s (1948) mathematical theory of signal transmission is of limited use in modeling human-human communication. First, the model assumes that the encoder function that the sender uses to convert a message into a signal is the inverse of the decoder function that the receiver uses to reconstruct the message from the signal. As an idealization, this is not descriptively adequate for human communication, which exhibits complex many-to-many mappings. Furthermore, in Shannon’s model, miscommunication is assumed to be caused solely by noise on the line, and thus can be compensated for by adding redundancy to the signal. With humans, although noise (e.g. in signal transmission) may indeed cause failures in communication, such failures more often arise from errors in the decoding of referring expressions or in the recognition of the communicative intention that motivated the signal (see Schegloff, 1987).

From a more linguistic perspective, Grice’s (1957) theory of meaning provides a very concise definition of what constitutes (intentional) communication, but is atheoretic as to how this process is accomplished. Research in this tradition encounters the daunting complexities, and potential infinite regress, associated with the recognition of mutual knowledge or common ground (Stalnaker, 1978; Clark and Marshall, 1981). An influential recent approach is the Interactive Alignment (IA) model of Pickering and Garrod (2004), which is a conceptual model of dialogue that relies on interactive alignment of the participants’ mental representations through the automatic and resource-poor process of priming. In successful communication, according to this account, the alignment of linguistic representations ultimately leads to alignment between participants’ situation models. Their approach is motivated by the intuition that full common ground processing is implausible given the speed and efficiency of typical dialogue.

We agree with this intuition but disagree about its implications for the modeling of dialogue. Notably, the IA model implicitly assumes even simpler encoder/decoder functions than Shannon’s model – namely, the identity function. This assumption does not do justice to the well-documented problem of the many-to-many mappings of utterances onto intentions and vice versa (see e.g. Levinson, 1983, 2000). For reasons of brevity, we refer to Krauss and Pardo.
(2004) for a more detailed discussion of some other important problems with the IA account.

The absence of models of human communication that are well-specified, implementable, and plausible from a linguistic standpoint is keenly felt. Such a model would facilitate both theoretical discussion and the generation of testable predictions about dialogue, and would enable the implementation of more human-like communication in artificial agents and robots. In what follows, we outline a mathematical model of communication that crucially relies upon the use of *shared conventions* to achieve efficiency, and that copes with the many-to-many mapping problem by using Bayesian updating. The model offers a general theoretical framework incorporating principled quantitative accounts of certain pervasive communicational phenomena, such as incrementality, repair, and gapless turn-taking.

In the following section we outline a more detailed motivation for the model, and present the problems that it should at least attempt to solve. We will then provide a formal specification of the model. Finally, we will briefly discuss its advantages, the predictions that it yields, and some interesting issues that arise from its formulation.

1. **The need for rapid intention recognition**

We aim to show first that hearers need some kind of rapid online heuristic in order to infer the communicative intentions of speakers. In our view, this follows from two major premises, which we will argue for in the following paragraphs. The first is that there is no time for intention recognition to take place in the gaps between conversational turns. The second is that the choice of appropriate conversational move relies crucially upon the correct identification of speaker intention.

In addition to the decoding of semantic meaning, Levinson (1995: 230) argues that intention recognition is crucial to the achievement of any kind of coordination between discourse participants. He emphasizes the difficulty of this operation, and in particular the matter of time pressure. As he observes, "since conversational response can routinely fall within 200 milliseconds of the prior utterance, if one modestly ascribes half of that delay to planning of the response, then that leaves only the other half for comprehension, including intention- or plan-recognition, of the prior utterance” (ibid.: 257). Moreover, the operation is formally intractable, as the same expression may typically encode many possible intentions, so the process of inferring the intention from the expression is not a simple lookup but must involve complex abductive inference. Subsequent research on conversational turn-taking and on utterance planning suggests that this timing problem is even more severe. Stivers et al. (2009) studied turn-taking in ten languages, and five of their sample exhibited a mean inter-turn latency of less than 200ms. For Japanese, this mean latency was 7ms. Meanwhile, research on utterance planning indicates that the time required to formulate an utterance is typically considerably longer than this (600ms elapsing between fixation on the referent and the start of utterance in Brown-Schmidt and Tanenhaus, 2006).

At the same time, intention recognition is crucial to the generation of appropriate responses. The same utterance can be used to perform multiple distinct functions: for instance, “It’s cold in here” might be a complaint, an accusation, an expression of sympathy, or a request (e.g. to close the window). Each of these would call for a different response from a competent interlocutor. Returning to the Stivers et al. (2009) study, it is evident that certain categories of response are privileged in their availability: significantly faster turn transitions were documented for answers than for non-answer responses, and confirmations were issued significantly faster than disconfirmations. In the case of Japanese, answers and confirmations both exhibited a negative mean latency (i.e. there was typically overlap with the preceding utterance). This suggests that some appropriate responses can be successfully planned long before the preceding utterance is completed.

Taken together, these observations seem to indicate that the identification of communicative intention necessarily takes place online during the relevant utterance. However, this does not prove that intention recognition takes place in parallel with the syntactic and semantic analysis of the utterance. Sacks, Schegloff and Jefferson (1974) argued that hearers anticipate the end of speakers’ turns and commence their utterance planning so as to avoid delays in turn-taking. De Ruiter, Mitterer and Enfield (2006) show that this process relies upon the lexicico-syntactic content of the speaker’s utterance. Moreover, Magyari and De Ruiter (2008) provide evidence that the accuracy of this anticipation is related to
Now the hearer’s ability to guess the words with which the speaker will end their conversational turn. If this is the case, it is theoretically possible that the hearer might guess the end of the turn, perform their semantic analysis on the basis of that guess, and only then proceed to compute the conversational intention, and that they still might be able to commence their own turn without leaving too long a gap in the conversation.

We consider it unlikely that such a sequential approach is tenable, for at least two reasons. First, as a matter of practicality, many utterances specify in more or less explicit terms their communicative intention: an utterance beginning “Could you” is highly likely to be a request, one beginning “I promise” is likely to be a promise, one beginning “I’m sorry” is likely to be an apology, and so on. It seems counterintuitive to suppose that this information is not available to the hearer to help in planning their response until the utterance is complete. Moreover, such information might usefully contribute to the hearer’s guess as to the turn completion itself (“Could you….please”). Second, the successful identification of communicative intention appears to be a prerequisite not only for the next conversational turn but also for appropriate listening behaviour. This observation follows a research tradition dating from Fries (1952), who proposed a typology of utterances including those “that have as responses continued attention” (p.49), and which elicit what were later labelled back-channel responses (Yngve, 1970) or continuers (Schegloff, 1982). Based on his corpus of telephone conversations, Fries identified a set of attentive responses, but these vary considerably in character, from “yes” and “good” to “oh my goodness” and “oh dear”. Use of the appropriate back-channel responses seems to require some on-line awareness of the speaker’s presumed intention: a back-channel response of “yes” might be appropriate to acknowledge a complaint or an explanation but would be potentially misleading in response to a request, especially if it is a request with which the hearer does not wish to comply.

Efficient exploitation of the available cues to speaker intention would serve to enable the hearer to give rapid, pertinent responses, without unnecessary delays. One further reason to suppose that this occurs in practice is the prevalence of behaviours that can be characterised by scripts (Schank and Abelson, 1977). Scripts specify the patterns of socially appropriate actions in particular circumstances, e.g. ordering food in a restaurant, where these patterns are constrained by the logical dependencies inherent in the situation. An utterance such as “The soup” may be ambiguous in principle between a request, reminder, complaint, etc., but if it is uttered by a diner who has just been seated, a waiter will naturally interpret it as a request. Moreover, this expectation appears to prefigure the fact of any specific utterance being made: a waiter might reasonably expect that if the diners name a dish, their utterance should be interpreted as a request for that dish to be served to them, in the absence of any specifications to the contrary (such as “What is the soup?”).

It would not only be inefficient but also odd if the rich set of expectations that participants carry about socially appropriate behaviours were not accessible until after the linguistic content of the utterance had been dealt with. Indeed, given that the interactive behaviours need not in general be linguistic, it would amount to saying that the linguistic channel has absolute priority over other communicative channels (such as gesture, posture, and para- and extra-linguistic behaviours that do not contribute to the truth-conditional meaning of the utterance), and that in the presence of speech, all other reasoning processes are placed on hiatus until the linguistic content is processed. This is perhaps a conceivable position, but we consider it implausible.

More generally, the reliance on prior knowledge and expectations may be crucial in circumventing the informational ‘bottleneck’ that is the speech system. Levinson (2000: 28) observes that human speech can convey maximally around 100 bits per second of information, which he considers “brutally slow”, whereas all other aspects of human linguistic competence can operate at a much faster rate. We agree with his suggestion that, consequently, “the design requirements are for a system that maximizes inference” (ibid.: 29). Because information arrives so slowly in the auditory channel, a system is favoured that uses this information incrementally and as soon as possible, rather than enduring the long wait for more input. For the same reason, the system should formulate rich initial hypotheses and look for support from the input data, rather than waiting for the data and only then attempting to reason from it. In the light of the findings on interactive behaviours, as summarised above, we consider that this argument applies to communicative intention recognition at least as
strongly as it does to local pragmatic enrichment processes.

In short, both experimental evidence and philosophical considerations support an account of intention recognition that makes use of prior knowledge along with new information, and that does so at the earliest opportunity, in order to generate the kind of rapid and relevant response that is characteristic of human communicative interactions. It is to the specification of such a model that we now turn.

2. **Specification of the model**

The AIRBUS model takes a signal as its input and calculates the corresponding intention. It has access to three forms of information: a convention database $C$, which specifies the probability of communicative intentions given certain signals; a likelihood database $L$, specifying the association between intentions and signals (i.e. the extent to which a particular intention activates particular signals, at a representational level); and a set of prior expectations $E$ as to the communicative intention. Entries in the convention database $C$ typically represent shared conventions about the function of particular signals, as discussed by De Ruiter and Enfield (2007). We would posit that $C$ and $L$ are also implicated in utterance formulation, a point we touch upon briefly in section 3 of this paper. The signals referred to by the model could in principle be both linguistic and non-linguistic – that is, this formalism could incorporate such behaviours as gesture, facial expression and so on – although the examples provided in section 2.2 will refer to linguistic signals.

Formally, the model can be specified as follows. Let $I$ denote the (finite) set of possible intentions. $E$ is the ordered set of perceived probabilities that each member of $I$ will be the communicative intention of the incoming utterance. We denote the probability that this intention will be the specific one $i \in I$ by the expression $E_i$, as follows.

$$\forall i \in I: E_i = P(\text{Intention} = i)$$

Thus for instance $E = (0.7, 0.3, 0, 0, 0\ldots)$ would denote the situation where there is a 0.7 probability of intention 1, a 0.3 probability of intention 2, and zero probability of all other intentions. Assuming that the set $I$ is exhaustive of possible intentions, the entries in $E$ sum to 1, and thus define a proper (discrete) probability distribution.

The convention database $C$ consists of entries for possible signals $j$, and specifies for each intention $i$ the probability that this is the appropriate intention given the signal (again as perceived by the hearer). Each signal has at most one corresponding entry in $C$. Let $C_{ij}$ denote the entry in $C$ for signal $j$, and $C_j$ denote the value in $C_j$ corresponding to the intention $i$. Then we have:

$$\forall i \in I: C_{ij} = P(\text{Intention} = i | \text{Signal} = j)$$

The likelihood database $L$ consists of a series of weights $w$, characterizing how strongly each intention activates each signal. We write $w_{ij}$ for the weight of the association between intention $i$ and signal $j$. This will be zero for signals that are not activated at all by a given intention.

The operation of the model consists of updating the prior probabilities $E$ in the light of a new incoming signal, taking into account the information in $C$ and $L$. We propose the following stages of update. Given a new signal $s$, the model examines whether there is an entry in $C$ corresponding to the signal $s$. If so, the probabilities in $C_i$ are merged with the probabilities in $E$, creating a set of revised probabilities $R$. Here we posit the use of a simple arithmetic mean, as follows; note that in this case the set of revised probabilities $R$ automatically sums to 1 and is thus a proper distribution.

$$R_i = \frac{E_i + C_{ij}}{2}$$

The set $R$ is then treated as a prior and is updated in the light of $L$. We apply Bayes’ Theorem:

$$P(i|s) = \frac{P(i)P(s|i)}{P(s)}$$

We wish to compute the probability of the intention given the signal. For any intention $i$, $R_i$ furnishes our prior probability that that intention is being communicated, $P(i)$ in this formula. We compute the ratio $P(s|i)/P(s)$ by using the information in the likelihood database $L$. Specifically, we consider this to be the strength of the activation of signal $s$ by intention $i$, expressed as a proportion of the total strength of the activation of $s$ by all possible intentions. This is calculated as follows.

$$\frac{P(s|i)}{P(s)} = \frac{w_{ij}}{\sum_{k \in I} w_{ik}}$$

For instance, suppose that a signal’s activation by intention 1 is 20, its activation by intention 2
is 60, and it is not activated by any other intentions. For this signal, if i is intention 1, the numerator of the above expression is 20, and its denominator is $20 + 60 = 80$. The expression thus evaluates to 0.25.

Note that, assuming that the weights are all non-negative, the above expression takes values in the range $[0, 1]$. However, the sum of the resulting $P(i|s)$ values will generally be less than 1. As a final step we therefore normalize the resulting values to turn them into a proper probability distribution over I.

2.1 Entropy of the distribution of intentions

The above formalism illustrates how the model maps communicative acts to probability distributions over the space of possible communicative intentions. Another way of looking at it is that the model manipulates a prior probability distribution by applying incoming information, turning it into a posterior distribution. From this perspective we can measure the success or the usefulness of a communicative act by considering the extent to which it reduces the hearer’s uncertainty as to the speaker’s intention. Following Shannon (1948), we can measure this by considering the entropy of the prior and posterior probability distributions over the possible intentions in I.

We posit that a successful communicative act generally results in lower entropy, in that there is more certainty as to the speaker’s intention after the act than there was before. We propose that the hearer commences the planning of a response when the hearer is sufficiently certain of the speaker’s intention – that is, when the entropy is low enough. Our model provides a convenient way to articulate this intuition: the two-step update process described above iterates throughout the speaker’s utterance, terminating when the entropy falls below a specified threshold. At this point the identified ‘most probable’ intention is assumed to be correct, and the hearer’s utterance planning proceeds on this basis. In this respect the model is inherently compatible with incremental processing, as the process of update is able to proceed (and potentially terminate) given partial information.

Separately, we can also appeal to the notion of relative entropy to propose an explanation of conversational repair mechanisms. Specifically, we propose that repair mechanisms are activated if there is too large a difference between prior and posterior distributions: that is, if the hearer’s understanding of the speaker’s intention is radically altered during the update process. A large difference would suggest disalignment between speaker and hearer, and the possible need for explicit repair negotiation.

We can measure this difference using Kullback-Leibler divergence, a standard measure of relative entropy, which quantifies the extent to which one distribution (here the posterior) is coded non-optimally if a coding scheme is used that is optimal for another distribution (here the prior). Very loosely, we might think of this as quantifying the intuition that highly unexpected new information will be difficult to comprehend within an existing world-view. The KL divergence from the posterior distribution P to the prior E is calculated as follows (taking $0 \log 0$ to be zero).

$$D_{KL}(P||E) = \sum_{i \in I} P(i) \log \frac{P(i)}{E(i)}$$

If the posterior distribution precisely matches the prior, the KL divergence evaluates to zero. Note that KL divergence is not symmetric. In particular, if $P(k) \approx 0$ for some intention k for which $E_k$ is considerably greater (i.e. a possible intention is ruled out on the basis of the communicative act), that does not suppose a large KL divergence. Contrastingly, if $P(k)$ is large for some k for which $E_k \approx 0$ (i.e. a possible intention becomes likely that was originally considered extremely unlikely), this does suppose a large KL divergence, because the optimal coding scheme for $P(r)$ will be one in which the intention k is not easily represented.

2.2 Worked example

We illustrate the operation of the model briefly here with an example in a restricted setting for which a corpus of interactions has been collected, namely a bar (Huth, Loth and De Ruiter, submitted). We consider a situation in which the bartender has made eye contact with a customer who is approaching the bar. For simplicity, we consider a space of just four communicative intentions: greeting, requesting (i.e. ordering a drink), asking a question, and making a statement. Based on Huth et al.’s data, we assign prior probabilities of (approximately) 0.15, 0.7, 0.15 and 0.001 respectively to these intentions\(^1\).

\(^1\) Formally these values must add to 1; for simplicity of presentation, when including 0.001 as a value here, we allow the other values to sum to 1.
Suppose first that the customer says “May I have a beer?” We would posit that the conventions activated by the initial components of this utterance specify that it is likely to be a request or a question. Let us assume that the entry in C for this utterance associates probabilities (0, 0.5, 0.5, 0) to the abovementioned intentions. In the first phase of update, these combine with the prior probabilities to give us a modified prior of (0.075, 0.6, 0.325, 0.0005).

We then consider the weights associated with the intentions. Here we might assume that this utterance is not typically caused by an intention to issue a greeting or to make a statement, but more likely arises from a request, and plausibly from a question. Let us assume that the corresponding weights are 1 for greeting, 50 for request, 10 for question, 1 for statement. The total weight is then 62, and in the second phase of update the prior probabilities must be multiplied by, respectively, 1/62, 50/62, 10/62 and 1/62. The resulting values are (to 3 d.p.) 0.001, 0.484, 0.052, and 0.000. Normalizing these, we find a posterior probability of 90.1% that the utterance was a request, versus 9.7% that it was a question. The resulting distribution has considerably lower entropy than the prior and indicates a high degree of confidence that the intention is a request.

Suppose instead that the customer says “Hello”. The prior probabilities are of course the same; let’s assume that the entry in C for the utterance specifies the probabilities (0.8, 0.01, 0.18, 0.01) and that the relevant weights in L are 100 for greeting, 10 for request, 30 for question, and 1 for statement. The first phase of update yields the probabilities (0.475, 0.355, 0.165, 0.0055); in the second phase of update these are multiplied by 100/141, 10/141, 30/141 and 1/141 to give (0.337, 0.025, 0.035, 0.000). Normalizing, we find a posterior probability of 84.9% that the utterance is a greeting. Again, the resulting distribution has lower entropy than the prior.

Finally, suppose that the customer first says “I’m sorry I’m late”. Here we might assume an entry in C along the lines of (0.25, 0.01, 0.01, 0.73), and relevant weights in L of 10 for greeting, 1 for request, 5 for question and 50 for statement. The first phase of update yields (0.2, 0.355, 0.08, 0.365); in the Bayesian update these are multiplied by 10/66, 1/66, 5/66 and 50/66 to give (0.030, 0.005, 0.006, 0.277). Normalizing, we find a posterior probability of 87.1% that the utterance is a statement. However, in this case, unlike the preceding examples, there is also a large KL divergence, reflecting the fact that a highly unlikely intention (according to the prior) is now the preferred interpretation. Under these conditions we would expect explicit repair to be initiated, thus according with our intuitions that the customer’s utterance is anomalous ‘out of the blue’ and requires explanation.

3. Discussion

We conclude by briefly sketching some of the implications of this model and the issues that must be addressed in its development. Generally, we argue that the model meets the desiderata outlined earlier in this paper, providing a rapid means to infer communicative intentions. The examples above illustrate the power of the two phases of operation that are posited. In the second and third examples, appealing to conventions does not suffice to obtain a clear ‘winning’ expectation. However, when the Bayesian step is also performed, the preference manifests itself. This reflects the additional strength offered by this decoding process, in that it uses the hearer’s knowledge about both directions of the relationship between signal and intention to draw pragmatic conclusions about the speaker’s intended meaning. Moreover, by its use of probability distributions rather than categorical rules, the model is able gracefully to handle improbable events. By equipping the model with a notion of relative entropy, we also have a means for predicting when this handling does break down to such an extent that explicit negotiation is required.

The account presented here considers the hearer’s task of intention recognition; however, its mechanisms could also provide the basis for a production model. Broadly, in such a model, the likelihood database L would suggest the possible utterances corresponding to the speaker’s (known) communicative intention. The convention database C would be used to monitor the proposed output and ensure that it did not give rise to unintended inferences about the speaker’s intention. However, our comprehension model is inherently and deliberately asymmetrical in the way it treats the information in C and L, and therefore the derivation of the corresponding production model is non-trivial. For reasons of space we do not attempt to specify such a model here.
Returning to comprehension, in this brief treatment we have necessarily left many issues open. We did not discuss how the likelihood database is to be populated: in practice this could be done based on actual exposure or through speakers and hearers simulating each other’s processing (De Ruiter et al., 2007, 2010; Noordzij et al., 2010). Likewise, we assumed in the examples that the convention database reflects actual experience on the part of the language user, but the mechanisms underlying this process are open to discussion. Another open question is which aspects of the utterance are listed in the convention database: that is, do the conventions relate to lexical items, syntactic categories (such as VP), or some other form of regular expression? A related theoretical issue is whether we consider this model to interact with the semantics, or consider the latter to be a Fodorian “module” that is cognitively impenetrable (Pylyshyn, 1984); we are not in a position yet to make any commitment either way, but clearly this has consequences for the predictions arising from our model. Finally, a very general question concerns the nature of the possible intentions themselves, an issue that has been explored from many directions. However, although we concede that the correct set of intentions must be posited in order to precisely simulate human behaviour, we would argue that the use of any plausible proxy set should be adequate in principle to achieve a close approximation to this behaviour.

In future work we aim to explore the capabilities of this model through a range of qualitative and quantitative tests. The model gives rise to testable predictions as to a wide range of behaviours. We could, for example, see whether the model can match humans’ attribution of communicative intentions across a range of discourse contexts. Under the assumption that a conversational turn may be planned when the entropy is sufficiently low, we could see whether the model correctly predicts turn latency across contexts. In conjunction with this, we could use the KL divergence measure to predict instance of repair, and consider whether that correctly identifies these occurrences in annotated corpora. Finally, we could implement the model within a robotic dialogue system, and explore whether it offers a perceived improvement to the existing behaviour. We feel that the model has considerable practical potential in providing enhanced artificial discourse capabilities, and that if this promise is borne out, it could also have substantial implications for the modelling of dialogic behaviour in natural language.

References


Kerstin Huth, Sebastian Loth, and J. P. de Ruiter. Submitted. How to order a beer: distilling computational models from natural data for social robotics.


