Deciding When To React To Incremental User Input In Human-Robot Interaction

Kathrin Janowski
Human Centered Multimedia
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany
kathrin.janowski@informatik.uni-augsburg.de

Elisabeth André
Human Centered Multimedia
Augsburg University
Universitätsstr. 6a
86159 Augsburg, Germany
andre@informatik.uni-augsburg.de

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems; H.5.2 [**Information Interfaces And Presentation**]: User Interfaces

General Terms
Algorithms, Human Factors

Keywords
incremental processing, uncertainty, utility, decision-theoretic approach

1. INTRODUCTION
As the technologies for interpreting and imitating human communication are improving, robots are becoming able to collaborate with humans using familiar modalities such as speech, gestures and gaze. Unfortunately, most of these interactions still lack the fluidity of natural human interaction, mostly because the robot has to wait for the user to finish their contribution before it can start to analyze it and choose the appropriate reaction.

Humans, in contrast, rely on subtle cues as well as their experience with established patterns for anticipating their partner's intentions or detecting misunderstandings. They constantly form hypotheses about the other person's mental state, adapting or discarding them whenever they receive more information. This in turn enables them to prepare appropriate reactions which can be performed as soon as or even before the partner has finished. Similarly, a computer system with this capability could speed up the interaction flow or intercept mistakes before they actually happen. The users' tendency to expect human-like reactions from robots, combined with the fact that current input technologies are still struggling to operate in real-time, makes this even more important [1].

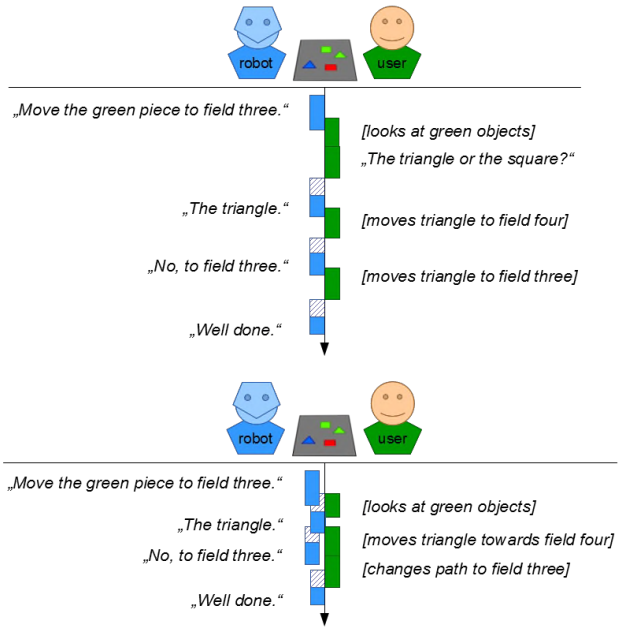

Figure 1: ***Upper image:*** **Typical interaction with the robot only interpreting finished inputs.** ***Lower image:*** **The same scenario augmented with incremental input processing.**

1.1 Motivating Example
Figure 1 shows a comparison between the typical human-robot interaction as observed in most dialogue systems nowadays and the more human-like version, illustrated with a simplified object placement task. In the upper image, the robot can only start to process the user's inputs after they have been completed, which results in a gap while it has to analyze the entire message at once. In addition to these delays, it is only able to detect the user's mistake after it has occurred.

In contrast, the lower image shows the same scenario with a robot capable of incremental processing. In analogy to the human perception described above, it starts analyzing the user's actions right from their beginning. There is still a delay before the robot arrives at a plausible interpretation, but this time the robot's reaction overlaps with the input.

Furthermore, by monitoring the movement of the triangle, the robot can already inform the user that it will land on the wrong field, allowing them to adjust the object's path seamlessly. Thanks to the early feedback, the user does not even need to ask for a clarification of the ambiguous instruction since the robot inferred their need for more information from their gaze pattern. The overall result of adding incremental processing to this scenario is a more fluid collaboration which avoids unnecessary pauses, communicative effort or mistakes.

1.2 Problem Description

Unfortunately, what appears to be plausible at the beginning of a sentence can turn out to be completely different after a crucial bit of information was added, so reacting before this point could cause more harm than good. Consider the user's gaze at the beginning of the interaction. After the robot has asked them to move "the green piece", the human is searching for an object which matches this description. Now there are several ways this interaction could play out. They might decide to ask the robot for a clarification, an effort which can be avoided if the robot adds the missing information on its own like depicted in the lower timeline. Consequently, clarifying the instruction now would allow the robot to save time and make the user more comfortable. Alternatively, the user might not ask such a question for various reasons. They might be about to find the correct piece on their own. In this case, the robot's help would not be necessary and if it intervened now, the robot might come across as impatient or patronizing which could upset the user despite the time saved. They could also decide to wait because they have no idea what to do (notice the parallel to the robot's decision problem), or even try moving a random piece, committing a mistake which would need to be undone before the interaction can continue. If the robot does not react in these cases, both would lose time and the user would perceive the robot as unhelpful.

The right moment for an action depends on a large number of factors, most of which cannot be observed directly but only assumed or inferred with a certain degree of confidence, like the user's understanding of the term "the green piece". A possible solution can be found in the way humans approach such decisions. Actions are usually taken if the potential benefits outweigh the risks and there is a high probability of success. For example, a listener might choose to interrupt the speaker when they believe the discussion was heading in the wrong direction and expect to improve its outcome by drawing attention to the important topic. On the other hand, they might be reluctant to do so while talking to somebody with higher authority, feeling that acting impatient towards them was more dangerous than talking about the wrong topic.

The following section gives an overview of related work on the reaction to uncertain or incomplete user inputs. Section 3 will then present a decision-theoretic approach for the question whether the robot should perform a given action at its current knowledge state. The paper ends with discussing some implications of this approach.

2. RELATED WORK

There are various approaches for the incremental interpretation of user inputs and the early resolution of ambiguities, often combining information from different modalities. For example, Cantrell et al. [1] narrow down the features for detecting objects in a robot's camera stream, such as color and position, whenever a new word is parsed from the user's speech. Kruijff et al. [5] described the use of modality-independent semantic representations for comparing the content obtained from different input channels, which makes it possible to reduce the number of hypotheses by unifying matching information. However, most of these works focus on finding the best interpretation and reacting to it as soon as possible, whereas few consider the circumstances under which the reaction should be delayed.

The first factor to consider before reacting is the degree of confidence in the current interpretation of the available data. Traum et al. [3, 7] have developed a multi-agent system which allows humans to train negotiation strategies in a virtual setting. They applied a machine-learning approach for determining the semantic content of the user's utterance so far, as well as predicting the content of the complete sentence and the expected confidence in this interpretation. This allows the virtual characters to either give early feedback during the user's turn [7] or to take over and finish an incomplete sentence when there is a pause in the user's speech and this confidence is high enough [3]. However, the authors note that this behavior would often be undesirable and, in the absence of a more sophisticated model, trigger it only when the pause was longer than a fixed threshold to avoid barging in on the user's turn.

As for the costs of interrupting the human, Horvitz and Apacible [4] modeled how users react to different forms of notifications while working on typical office tasks. They trained a Bayesian network for predicting the cost of interrupting the user in the current situation or the near future. Such a network models the causal relationships between certain observations as their conditional probabilities. So-called influence diagrams extend this model further by adding utility nodes which describe the costs or benefits associated with these observations. Since a Bayesian network can infer probability distributions for all of its nodes from values observed for some of them, an influence diagram can calculate the expected utility of a decision by adding up the costs and benefits for all its possible outcomes, weighted by their respective probabilities. [6]

Conati [2] describes the use of a Dynamic Decision Network, which works on the same mathematical principles, for modeling the dependencies between the user's perceivable emotional expressions, their affective state and its causes. Since the latter cannot be observed directly, they need to be inferred from different, possibly incomplete sensor inputs. Utility functions included in this network then allow a virtual assistant to choose the behavior which is most likely to be appreciated by the user, or decide to delay an action if it is not desirable. The problem of incomplete inputs is very similar to that of an incremental system trying to interpret a user's message before it was complete, which makes this approach useful for both cases.

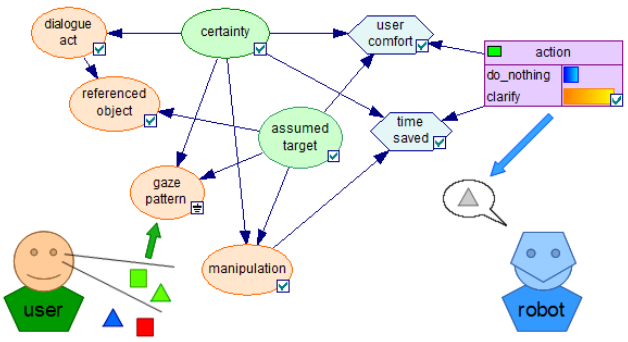

Figure 2: A draft for an influence diagram which predicts the consequences of a clarification action.

3. SOLUTION APPROACH

Remember the scenario from figure 1 and the possible outcomes for clarifying the instruction while the user is looking for the referenced object. Figure 2 shows a draft for solving this problem with an influence diagram which contains probability nodes for the user's mental state and behavior, utility nodes for the user's level of comfort and the time which can be saved, and finally a decision node with the two options "do nothing" and "clarify" for the robot's behavior.

The user's mental state consists of their assumption regarding the object they should move and the degree of certainty for this assumption. Both of these factors influence their behavior, which can consist of different gaze patterns, asking different types of question about any of the available objects or physically manipulating one of these objects. The user's level of comfort depends on whether the robot's help is required, which is the case when the user misunderstood the instruction or is uncertain about the target object. The amount of time which can be saved by the robot's action depends on the user's manipulation action, which either completes the task by placing the object in the correct location or hinders it in case of an incorrect placement. It also depends on the user's level of certainty which determines the likelihood that the user will hesitate before moving one of the objects.

After giving the instruction, the robot monitors whether the user reacts as expected. When the robot's sensors observe a change in the user's behavior, in this case a gaze shift towards several green objects, it updates the network with this data. In order to reason about the consequences of its action, the robot first needs to choose at least one action suitable for the situation it appears to be in. In this case, a fixed rule in its behavior manager suggests to clarify the previous sentence if the user does not react by moving the correct piece. Now the robot needs to determine the possible effects of performing or avoiding this action, not only on the most likely situation but on any other situation which could explain the sensor readings. So it calculates the time that can be saved or lost, as well as the impact on the user's mood, depending on the user's possible mental states and the object manipulation that might follow. These values are stored in the utility nodes. Finally, the network is updated in order to obtain the missing probabilities given the user's gaze behavior and calculate the expected utility for the two reaction options "clarify" and "do nothing". These utilities then tell the robot which option has the preferable outcome and consequently, it will clarify its instruction if the utility for this action is higher than that for doing nothing and waiting for the next sensor input.

4. DISCUSSION

Models following the principle of a Bayesian network are very flexible with regards to the available data. The robot could base its decision on any of the behaviors related to the user's mental state, for example on the content of a partial sentence in addition to or instead of the gaze pattern. Furthermore, the evidence could be based on low-level features of the behavior (as mentioned in [2]) as well as higher-level semantic interpretations obtained from a specialized input processing module. For instance, the model could reason that a particular hand movement is part of various gestures related to different intentions, leading to different interpretations of the user's mindset, or rely on a classifier tailored to gesture analysis for determining which gesture the user is most likely performing. The former is an example for observations made near the beginning of the input whereas the latter may only be available after a significant portion of the user's input has been processed, so the model would be prepared to deal with the input at any stage of the interpretation process.

Furthermore, as more reliable data is becoming available, the probability for observing one particular situation increases. Because the expected benefits of an outcome scale with its likelihood, this in turn increases the chance that the robot will perform the chosen action, reflecting the intuitive idea that it should act when it is confident enough and wait for more data otherwise. Likewise, when the new data contradicts the previous hypothesis, the probability and utility decrease and so the robot will keep waiting.

One more advantage of this approach is its ability to model adaptive priorities for different behaviors. For example, when the user is focusing on the correct object, the robot might confirm this assumption with a verbal comment, a short nod or both. Compared to speech, an unobtrusive head gesture might have a comparatively low benefit when the user requires feedback, but also far lower costs if this help is not wanted. Therefore, it might produce a higher expected utility and be chosen earlier when the user's attention focus is still ambiguous whereas the comment would follow at a later point in time. On the other hand, when a mistake is likely to occur, the expected utility for speech would rise more quickly since the benefit of avoiding the mistake would exceed the cost of being obtrusive. Consequently, the robot could follow the same behavior policies throughout the interaction.

A possible drawback of using the same mechanisms for all the robot's actions might be the overhead of querying the network for every single input event, some of which could be handled with simpler rules. In order to keep the architecture consistent, it may be worthwhile to find strategies for handling different levels of complexity with the same model. For example, substructures could be deactivated depending on the requirements of certain interaction contexts or marked as optional in the definition of individual behaviors.

5. REFERENCES

[1] R. Cantrell, E. Krause, M. Scheutz, M. Zillich, and E. Potapova. Incremental referent grounding with nlp-biased visual search. In *Proceedings of AAAI 2012 Workshop on Grounding Language for Physical Systems*, July 2012.

[2] C. Conati. Virtual butler: What can we learn from adaptive user interfaces? In R. Trappl, editor, *Your Virtual Butler*, volume 7407 of *Lecture Notes in Computer Science*, pages 29–41. Springer Berlin Heidelberg, 2013.

[3] D. DeVault, K. Sagae, and D. Traum. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1):143–170, 2011.

[4] E. Horvitz and J. Apacible. Learning and reasoning about interruption. In *Proceedings of the 5th international conference on Multimodal interfaces*, ICMI '03, pages 20–27, New York, NY, USA, 2003. ACM.

[5] G.-J. M. Kruijff, P. Lison, T. Benjamin, H. Jacobsson, and N. Hawes. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*, pages 55–64, December 2007.

[6] R. E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ 07458, 2004.

[7] D. Traum, D. DeVault, J. Lee, Z. Wang, and S. Marsella. Incremental dialogue understanding and feedback for multiparty, multimodal conversation. In Y. Nakano, M. Neff, A. Paiva, and M. Walker, editors, *Intelligent Virtual Agents*, volume 7502 of *Lecture Notes in Computer Science*, pages 275–288. Springer Berlin Heidelberg, 2012.