Expanding the Horizons of Natural Language Interfaces

Phil Hayes
Computer Science Department, Carnegie-Mellon University
Pittsburgh, PA 15213, USA

Abstract

Current natural language interfaces have concentrated largely on determining the literal "meaning" of input from their users. While such decoding is an essential underpinning, much recent work suggests that natural language interfaces will never appear cooperative or graceful unless they also incorporate numerous non-literal aspects of communication, such as robust communication procedures.

This paper defends that view, but claims that direct imitation of human performance is not the best way to implement many of these non-literal aspects of communication; that the new technology of powerful personal computers with integral graphics displays offers techniques superior to those of humans for these aspects, while still satisfying human communication needs. The paper proposes interfaces based on a judicious mixture of these techniques and the still valuable methods of more traditional natural language interfaces.

1. Introduction

Most work so far on natural language communication between man and machine has dealt with its literal aspects. That is, natural language interfaces have implicitly adopted the position that their user's input encodes a request for information or action, and that their job is to decode the request, retrieve the information, or perform the action, and provide appropriate output back to the user. This is essentially what Thomas [24] calls the Encoding-Decoding model of conversation.

While literal interpretation is a basic underpinning of communication, much recent work in artificial intelligence, linguistics, and related fields has shown that it is far from the whole story in human communication. For example, appropriate interpretation of an utterance depends on assumptions about the speaker's intentions, and conversely, the speaker's goals influence what is said (Hobbs [13], Thomas [24]). People often make mistakes in speaking and listening, and so have evolved conventions for effecting repairs (Schegloff et al. [20]). There must also be a way of regulating the turns of participants in a conversation (Sacks et al. [19]). This is just a sampling of what we will collectively call *non literal aspects of communication*.

The primary reason for using natural language in man-machine communication is to allow the user to express himself naturally, and without having to learn a special language. However, it is becoming clear that providing for natural expression means dealing with the non-literal as well as the literal aspects of communication; that the ability to interpret natural language literally does not in itself give a man-machine interface the ability to communicate naturally. Some work on incorporating these non-literal aspects of communication into man-machine interfaces has already begun ([6, 8, 9, 15, 21, 25]).

The position I wish to stress in this paper is that natural language interfaces will never perform acceptably unless they deal with the non-literal as well as the literal aspects of communication; that without the non-literal aspects, they will always appear uncooperative, inflexible, unfriendly, and generally stupid to their users, leading to irritation, frustration, and an unwillingness to continue to be a user.

This position is coming to be held fairly widely. However, I wish to go further and suggest that, in building non-literal aspects of communication into natural-language interfaces, we should aim for the most effective type of communication rather than insisting that the interface model human performance as exactly as possible. I believe that these two aims are not necessarily the same, especially given certain new technological trends discussed below.

Most attempts to incorporate non-literal aspects of communication into natural language interfaces have attempted to model human performance as closely as possible. The typical mode of communication in such an interface, in which system and user type alternately on a single scroll of paper (or scrolled display screen), has been used as an analogy to normal spoken human conversation in which communication takes place over a similar half-duplex channel, i.e. a channel that only one party at a time can use without danger of confusion.

Technology is outdating this model. The nascent generation of powerful personal computers (e.g. the ALTO [23] or PERQ [18]) equipped with high-resolution bit-map graphics display screens and pointing devices allow the rapid display of large quantities of information and the maintenance of several independent communication channels for both output (division of the screen into independent windows, highlighting, and other graphics techniques), and input (direction of keyboard input to different windows, pointing input). I believe that this new technology can provide highly effective, natural language-based, communication between man and machine, but only if the half-duplex style of interaction described above is dropped. Rather than trying to imitate human conversation directly, it will be more fruitful to use the capabilities of this new technology, which in some respects exceed those possessed by humans, to achieve the same ends as the non-literal aspects of normal human conversation. Work by, for instance, Carey [3] and Hiltz [12] shows how adaptable people are to new communication situations, and there is every reason to believe that people will adapt well to an interaction in which their communication *needs* are satisfied, even if they are satisfied in a different way than in ordinary human conversation.

In the remainder of the paper I will sketch some human communication needs, and go on to suggest how they can be satisfied using the technology outlined above.

2. Non-Literal Aspects of Communication

In this section we will discuss four human communication needs and the non-literal aspects of communication they have given rise to:

- non-grammatical utterance recognition
- contextually determined interpretation
- robust communication procedures
- channel sharing

The account here is based in part on work reported more fully in [8, 9].

Humans must deal with **non-grammatical utterances** in conversation simply because people produce them all the time. They arise from various sources: people may leave out or swallow words; they may start to say one thing, stop in the middle, and substitute something else; they may interrupt themselves to correct something they have just said; or they may simply make errors of tense, agreement, or vocabulary. For a combination of these and other reasons, it is very rare to see three consecutive grammatical sentences in ordinary conversation.

Despite the ubiquity of ungrammaticality, it has received very little attention in the literature or from the implementers of natural-language interfaces. Exceptions include PARRY [17], COOP [14], and interfaces produced by the LIFER [11] system. Additional work on parsing ungrammatical input has been done by Weischedel and Black [25], and

Kwasny and Sandheimer [15]. As part of a larger project on user interfaces [1], we (Hayes and Mouradian [7]) have also developed a parser capable of dealing flexibly with many forms of ungrammaticality.

Perhaps part of the reason that flexibility in parsing has received so little attention in work on natural language interfaces is that the input is typed, and so the parsers used have been derived from those used to parse written prose. Speech parsers (see for example [10] or [26]) have always been much more flexible. Prose is normally quite grammatical simply because the writer has had time to make it grammatical. The typed input to a computer system is produced in "real time" and is therefore much more likely to contain errors or other ungrammaticalities.

The listener at any given turn in a conversation does not merely decode or extract the inherent "meaning" from what the speaker said. Instead, he interprets the speaker's utterance in the light of the total available context (see for example, Hobbs [13], Thomas [24], or Wynn [27]). In cooperative dialogues, and computer interfaces normally operate in a cooperative situation, this **contextually determined interpretation** allows the participants considerable economies in what they say, substituting pronouns or other anaphoric forms for more complete descriptions, not explicitly requesting actions or information that they really desire, omitting participants from descriptions of events, and leaving unsaid other information that will be "obvious" to the listener because of the context shared by speaker and listener. In less cooperative situations, the listener's interpretations may be other than the speaker intends, and speakers may compensate for such distortions in the way they construct their utterances.

While these problems have been studied extensively in more abstract natural language research (for just a few examples see [4, 5, 16]), little attention has been paid to them in more applied language work. The work of Grosz [6] and Sidner [21] on focus of attention and its relation to anaphora and ellipsis stand out here, along with work done in the COOP [14] system on checking the presuppositions of questions with a negative answer. In general, contextual interpretation covers most of the work in natural language processing, and subsumes numerous currently intractable problems. It is only tractable in natural language interfaces because of the tight constraints provided by the highly restricted worlds in which they operate.

Just as in any other communication across a noisy channel, there is always a basic question in human conversation of whether the listener has received the speaker's utterance correctly. Humans have evolved **robust communication** conventions for performing such checks with considerable, though not complete, reliability, and for correcting errors when they occur (see Schegloff [20]). Such conventions include: the speaker assuming an utterance has been heard correctly unless the reply contradicts this assumption or there is no reply at all; the speaker trying to correct his own errors himself; the listener incorporating his assumptions about a doubtful utterance into his reply; the listener asking explicitly for clarification when he is sufficiently unsure.

This area of robust communication is perhaps the non-literal aspect of communication most neglected in natural language work. Just a few systems such as LIFER [11] and COOP [14] have paid even minimal attention to it. Interestingly, it is perhaps the area in which the new technology mentioned above has the most to offer as we shall see.

Finally, the spoken part of a human conversation takes place over what is essentially a single shared channel. In other words, if more than one person talks at once, no one can understand anything anyone else is saying. There are marginal exceptions to this, but by and large reasonable conversation can only be conducted if just one person speaks at a time. Thus people have evolved conventions for **channel sharing** [19], so that people can take turns to speak. Interestingly, if people are put in new communication situations in which the standard turn-taking conventions do not work well, they appear quite able to evolve new conventions [3].

As noted earlier, computer interfaces have sidestepped this problem by making the interaction take place over a half-duplex channel somewhat analogous to the half-duplex channel inherent in speech, i.e. alternate turns at typing on a scroll of paper (or scrolled display screen). However, rather than providing flexible conventions for changing turns, such interfaces typically brook no interruptions while they are typing, and then when they are finished insist that the user type a complete input with no feedback (apart from character echoing), at which point the system then takes over the channel again.

In the next section we will examine how the new generation of interface technology can help with some of the problems we have raised.

3. Incorporating Non-Literal Aspects of Communication into User Interfaces

If computer interfaces are ever to become cooperative and natural to use, they must incorporate non-literal aspects of communication. My main point in this section is that there is no reason they should incorporate them in a way directly imitative of humans; so long as they are incorporated in a way that humans are comfortable with, direct imitation is not necessary. Indeed, direct imitation is unlikely to produce satisfactory interaction. Given the present state of natural language processing and artificial intelligence in general, there is no prospect in the forseeable future that interfaces will be able to emulate human performance, since this depends so much on bringing to bear larger quantities of knowledge than current AI techniques are able to handle. Partial success in such emulation is only likely to raise false expectations in the mind of the user, and when these expectations are inevitably crushed, frustration will result. However, I believe that by making use of some of the new technology mentioned earlier, interfaces can provide very adequate substitutes for human techniques for non-literal aspects of communication; substitutes that capitalize on capabilities of computers that are not possessed by humans, but that nevertheless will result in interaction that feels very natural to a human.

Before giving some examples, let us review the kind of hardware I am assuming. The key item is a bit-map graphics display capable of being filled with information very quickly. The screen can be divided into independent windows to which the system can direct different streams of output independently. Windows can be moved around on the screen, overlapped, and popped out from under a pile of other windows. The user has a pointing device with which he can position a cursor to arbitrary points on the screen, plus, of course, a traditional keyboard. Such hardware exists now and will become increasingly available as powerful personal computers such as the PERQ [18] or LISP machine [2] come onto the market and start to decrease in price. The examples of the use of such hardware which follow are drawn in part from our current experiments in user interface research [1, 7] on similar hardware.

Perhaps the aspect of communication that can receive the most benefit from this type of hardware is robust communication. Suppose the user types a non-grammatical input to the system which the system's flexible parser is able to recognize if, say, it inserts a word and makes a spelling correction. Going by human convention the system would either have to ask the user to confirm explicitly if its correction was correct, to cleverly incorporate its assumption into its next output, or just to assume the correction without comment. Our hypothetical system has another option: it can alter what the user just typed (possibly highlighting the words that it changed). This achieves the same effect as the second option above, but substitutes a technological trick for human intelligence.

Again, if the user names a person, say "Smith", in a context where the system knows about several Smiths with different first names, the human options are either to incorporate a list of the names into a sentence (which becomes unwieldy when there are many more than three alternatives) or to ask for the first name without giving alternatives. A third alternative, possible only in this new technology, is to set up a window on the screen with an initial piece of text followed by a list of alternatives (twenty can be handled quite naturally this way). The user is then free to point at the alternative he intends, a much simpler and more natural alternative than typing the name, although there is no reason why this input mode should not be available as well in case the user prefers it.

As mentioned in the previous section, contextually based interpretation is important in human conversation because of the economies of expression it allows. There is no need for such economy in an interface's output, but the human tendency to economy in this matter is something that technology cannot change. The general problem of keeping track of focus of attention in a conversation is a difficult one (see, for example, Grosz [6] and Sidner [22]), but the type of interface we are discussing can at least provide a helpful framework in which the current focus of attention can be made explicit. Different foci of attention can be associated with different windows on the screen, and the system can indicate what it thinks is the current focus of attention by, say, making the border of the corresponding window different from all the rest. Suppose in the previous example that at the time the system displays the alternative Smiths, the user decides that he needs some other information before he can make a selection. He might ask for this information in a typed request, at which point the system would set up a new window, make it the focused window, and display the requested information in it. At this point, the user could input requests to refine the new information, and any anaphora or ellipsis he used would be handled in the appropriate context.

Representing contexts explicitly with an indication of what the system thinks is the current one can also prevent confusion. The system should try to follow a user's shifts of focus automatically, as in the above example. However, we cannot expect a system of limited understanding always to track focus shifts correctly, and so it is necessary for the system to give explicit feedback on what it thinks the shift was. Naturally, this implies that the user should be able to change focus explicitly as well as implicitly (probably by pointing to the appropriate window).

Explicit representation of foci can also be used to bolster a human's limited ability to keep track of several independent contexts. In the example above, it would not have been hard for the user to remember why he asked for the additional information and to return and make the selection after he had received that information. With many more than two contexts, however, people quickly lose track of where they are and what they are doing. Explicit representation of all the possibly active tasks or contexts can help a user keep things straight.

All the examples of how sophisticated interface hardware can help provide non-literal aspects of communication have depended on the ability of the underlying system to produce possibly large volumes of output rapidly at arbitrary points on the screen. In effect, this allows the system multiple output channels independent of the user's typed input, which can still be echoed even while the system is producing other output. Potentially, this frees interaction over such an interface from any turn-taking discipline. In practice, some will probably be needed to avoid confusing the user with too many things going on at once, but it can probably be looser than that found in human conversations.

As a final point, I should stress that natural language capability is still extremely valuable for such an interface. While pointing input is extremely fast and natural when the object or operation that the user wishes to identify is on the screen, it obviously cannot be used when the information is not there. Hierarchical menu systems, in which the selection of one item in a menu results in the display of another more detailed menu, can deal with this problem to some extent, but the descriptive power and conceptual operators of natural language (or an artificial language with similar characteristics) provide greater flexibility and range of expression. If the range of options is large, but well discriminated, it is often easier to specify a selection by description than by pointing, no matter how cleverly the options are organized.

4. Conclusion

In this paper, I have taken the position that natural language interfaces to computer systems will never be truly natural until they include non-literal as well as literal aspects of communication. Further, I claimed that in the light of the new technology of powerful personal computers with integral graphics displays, the best way to incorporate these non-literal aspects was not to imitate human conversational patterns as closely as possible, but to use the technology in innovative ways to perform the same function as the non-literal aspects of communication found in human conversation.

In any case, I believe the old-style natural language interfaces in which the user and system take turns to type on a single scroll of paper (or scrolled display screen) are doomed. The new technology can be used, in ways similar to those outlined above, to provide very convenient and attractive interfaces that do not deal with natural language. The advantages of this type of interface will so dominate those associated with the old-style natural language interfaces that continued work in that area will become of academic interest only.

That is the challenge posed by the new technology for natural language interfaces, but it also holds a promise. The promise is that a combination of natural language techniques with the new technology will result in interfaces that will be truly natural, flexible, and graceful in their interaction. The multiple channels of information flow provided by the new technology can be used to circumvent many of the areas where it is very hard to give computers the intelligence and knowledge to perform as well as humans. In short, the way forward for natural language interfaces is not to strive for closer, but still highly imperfect, imitation of human behaviour, but to combine the strengths of the new technology with the great human ability to adapt to communication environments which are novel but adequate for their needs.

References

1. Ball, J. E. and Hayes, P. J. Representation of Task-Independent Knowledge in a Gracefully Interacting User Interface. Tech. Rept. , Carnegie-Mellon University Computer Science Department, 1980.

2. Bawden, A, et al. Lisp Machine Project Report. AIM 444, MIT AI Lab, Cambridge, Mass., August, 1977.

3. Carey, J. "A Primer on Interactive Television." *J. University Film Assoc. XXX*, 2 (1978), 35-39.

4. Charniak, E. C. Toward a Model of Children's Story Comprehension. TR-266, MIT AI Lab, Cambridge, Mass., 1972.

5. Cullingford, R. *Script Application: Computer Understanding of Newspaper Stories.* Ph.D. Th., Computer Science Dept., Yale University, 1978.

6. Grosz, B. J. The Representation and Use of Focus in a System for Understanding Dialogues. Proc. Fifth Int. Jt. Conf. on Artificial Intelligence, MIT, 1977, pp. 67-76.

7. Hayes, P. J. and Mouradian, G. V. Flexible Parsing. Proc. of 18th Annual Meeting of the Assoc. for Comput. Ling., Philadelphia, June, 1980.

8. Hayes, P. J., and Reddy, R. Graceful Interaction in Man-Machine Communication. Proc. Sixth Int. Jt. Conf. on Artificial Intelligence, Tokyo, 1979, pp. 372-374.

9. Hayes, P. J., and Reddy, R. An Anatomy of Graceful Interaction in Man-Machine Communication. Tech. report, Computer Science Department, Carnegie-Mellon University, 1979.

10. Hayes-Roth, F., Erman, L. D., Fox, M., and Mostow, D. J. Syntactic Processing in HEARSAY-II. Speech Understanding Systems. Summary of Results of the Five-Year Research Effort at Carnegie-Mellon University, Carnegie-Mellon University Computer Science Department, 1976.

11. Hendrix, G. G. Human Engineering for Applied Natural Language Processing. Proc. Fifth Int. Jt. Conf. on Artificial Intelligence, MIT, 1977, pp. 183-191.

12. Hiltz, S. R., Johnson, K., Aronovitch, C., and Turoff, M. Face to Face vs. Computerized Conferences: A Controlled Experiment. unpublished mss.

13. Hobbs, J. R. Conversation as Planned Behavior. Technical Note 203, Artificial Intelligence Center, SRI International, Menlo Park, Ca., 1979.

14. Kaplan, S. J. *Cooperative Responses from a Portable Natural Language Data Base Query System*. Ph.D. Th., Dept. of Computer and Information Science, University of Pennsylvania, Philadelphia, 1979.

15. Kwasny, S. C. and Sondheimer, N. K. Ungrammaticality and Extra-Grammaticality in Natural Language Understanding Systems. Proc. of 17th Annual Meeting of the Assoc. for Comput. Ling., La Jolla, Ca., August, 1979, pp. 19-23.

16. Levin, J. A., and Moore, J. A. "Dialogue Games: Meta-Communication Structures for Natural Language Understanding." *Cognitive Science 1*, 4 (1977), 395-420.

17. Parkison, R. C., Colby, K. M., and Faught, W. S. "Conversational Language Comprehension Using Integrated Pattern-Matching and Parsing." *Artificial Intelligence 9* (1977), 111-134.

18. PERQ. Three Rivers Computer Corp., 160 N. Craig St., Pittsburgh, PA 15213. .

19. Sacks, H., Schegloff, E. A., and Jefferson, G. "A Simplest Semantics for the Organization of Turn-Taking for Conversation." *Language 50*, 4 (1974), 696-735.

20. Schegloff, E. A., Jefferson, G., and Sacks, H. "The Preference for Self-Correction in the Organization of Repair in Conversation." *Language 53*, 2 (1977), 361-382.

21. Sidner, C. L. A Progress Report on the Discourse and Reference Components of PAL. A. I. Memo. 468, MIT A. I. Lab., 1978.

22. Sidner, C. L. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. TR 537, MIT AI Lab, Cambridge, Mass., 1979.

23. Thacker, C.P., McCreight, E.M., Lampson, B.W., Sproull, R.F., and Boggs, D.R. Alto: A personal computer. In *Computer Structures: Readings and Examples*, McGraw-Hill, 1980. Edited by D. Siewiorek, C.G. Bell, and A. Newell, second edition, in press.

24. Thomas, J. C. "A Design-Interpretation of Natural English with Applications to Man-Computer Interaction." *Int. J. Man-Machine Studies 10* (1978), 651-668.

25. Weischedel, R. M. and Black, J. Responding to Potentially Unparseable Sentences. Tech. Rept. 79/3, Dept. of Computer and Information Sciences, University of Delaware, 1979.

26. Woods, W. A., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., and Zue, V. Speech Understanding Systems - Final Technical Report. Tech. Rept. 3438, Bolt, Beranek, and Newman, Inc., 1976.