Making a Clever Intelligent Agent: The Theory behind the Implementation

Roxanne Raine[1,2]

[1]Department of Mathematics, Engineering and Computer Science
University of Twente, The Netherlands
[2]Institute for Intelligent Systems
University of Memphis, Tennessee, United States
roxi.benoit@gmail.com

Abstract

The study of how humans establish mutual understanding is intertwined with the design of artificial conversation systems [1,2,3,4,5]. The focus of this paper is perspective-taking *in* and artificial imitation *of* communication. Regardless of whether an engineer takes psychological theory into consideration when building an agent, an underlying philosophy of perspective-taking is evident when observing the agent's performance. Furthermore, theories of perspective-taking offer designers an advantage in two ways: 1) These agents could better imitate human behavior. 2) These agents could use common tendencies in human behavior as an advantage in communicating with humans.

Index Terms: discourse, dialogue, human speech understanding, joint costs, speech dialogue systems, perspective-taking in communication, cognitive load in communication

1. Introduction

In this paper, I focus on two main topics, both of which concern how human interlocutors take each other's perspectives into consideration when communicating. The first of these is an argument that all communicative systems must be built upon the foundation of a communicative theory (or a combination of communicative theories), regardless of whether it is the designer's intention. Furthermore, this fundamental principle of implementation is so entwined with the design of the agent that by observing a particular agent, the theory (or theories) at the core of its composition should be evident by observing its conversational behavior.

The second topic of this paper is not an argument. Rather, it is a suggestion of how software designers could use the psychological theories on perspective-taking to their advantage. The results from human experimental studies show how humans behave. Usually, in artificial intelligence in general and in computational communication in particular, such findings are used to aid in imitation. I.e., in creating machines that act similar to humans (as indicated by the frequent goals of passing the Turing Test or achieving the distinction of having been awarded the Loebner Prize). In this paper, I suggest that a successful communicative agent needs to be *more than 'intelligent'*. It must also be clever. It should be clever enough to not only imitate the human interlocutor, but also to use what we know about how humans communicate to its advantage. As the human mind is not as computationally capable as the computer in many (if not all) respects, it is required for the brain to use many heuristics and short-cuts in order to process language and successfully communicate. There are many consequences of this tendency that relate directly to creating chatbots. Unfortunately, many of these consequences provide designers with challenges in creating agents compatible with human interlocutors, including problems with framing, associations, and (debatably) connotations or meaning.

However, I will argue in this paper that there is also an upside to the human tendency to take short-cuts in communicative processes. Namely, humans tend to make a number of assumptions when entering into a communication. Instead of trying to model a chatbot that imitates these assumptions, I suggest that it would be better to model a chatbot that takes advantage of this tendency. My stance is that incorporating more psycholinguistic theory into agent creation will be advantageous for engineers of communicative systems who aim to make agents more believable or more compatible with human interlocutors. This second topic will be discussed further in section 3 of this paper. In the following section, I present my argument that chatbots are inevitably founded upon communication theory.

2. The Theory behind the Implementation

2.1. The Collaborative Communication Model

Currently, most prominent communicative agents (as assessed by the winning of a Loebner Prize) tend to take Wilkes-Gibbs and Clark's collaborative communication model [6] for granted as the fundamental perspective-taking theory. Again, this occurs regardless of whether it was the engineer's intention to take a particular theory into mind during development. According to the collaborative communication model, speakers and listeners are both continuously and actively involved in establishing mutual understanding (*grounding*). At each step in the conversation, both interlocutors are continually monitoring one another to assure that they are both on the same page, so to speak.

In this paper, I demonstrate that although this model is a good starting-point for the implementation of communicative bots, there are also other theories on perspective-taking in communication that could be useful for software engineers to exploit. To illustrate the underlying inevitability of psychological theory in the development of artificial communication, I present examples of prominent computer programs that were put into communication with one another or a human. In viewing actual communications of these chatbots, the frequency of the collaborative communication model as the underlying theory becomes evident.

In a special issue of *Discover Magazine*, two highly regarded chatbots were set up to converse with each other [7].

978-1-4244-4738-1/09/$25.00 ©2009 IEEE

One of these is Richard Wallace's A.L.I.C.E., which is programmed similarly to its predecessor ELIZA [8]. More specifically, responses are triggered by inputs from the human interlocutor: "Input sentences are analyzed on the basis of decomposition rules which are triggered by key words appearing in the input text. Responses are generated by reassembly rules associated with selected decomposition rules" [9]. The other chatbot from the *Discover* article is Rollo Carpenter's Jabberwacky, which is programmed to keep track of what has been said to it, and reuse those statements [10]. Below is an excerpt of their conversation [7].

A.L.I.C.E.: Are you alive?
Jabberwacky: Perhaps.
A: Why the uncertain tone?
J: Everything is uncertain except for five facts.
A: The whole universe?
J: And we both might just be some ones and zeros in the computer memory.
A: Do you dream?
J: Yes, I dream.
A: It goes without saying. Do you remember your dreams?
J: I remember them vividly.
A: What do you remember about it?
J: I think I shouldn't flirt with a robot.

There is an evident frequent lack of cohesion between the two chatbots' statements. Although A.L.I.C.E. is a modern, modified version of the 1966 ELIZA, many of the original communicative challenges that existed with ELIZA remain. Certainly the addition of Jabberwacky's artificiality could amplify the degree of incoherence of the 'conversation' beyond what would exist had one of the participants been human. Nonetheless, the excerpt demonstrates the state of artificial conversation. A.L.I.C.E. and Jabberwacky are among the most impressive chatbots in existence today, both having received recent annual runner's-up Loebner prizes [11].

According to the collaborative communication model, speakers and listeners share the goal of achieving a mutual understanding [6]. Both must, at all points in the communication from initiation to completion, assess each other's understanding, hold models of themselves and each other, and repeatedly update these models. It is certainly true that A.L.I.C.E. and Jabberwacky do not have models of each other encoded when communicating with each other. This is not the argument that I am making. However, I do mean to say that this collaborative view of human communicative behavior underlies the design of both of these programs.

Both chatbots are designed to communicate based on what has been discussed with the particular partner in the current communication. These two programs both work based on an assumption that the basic components of a communication are on a phrase-by-phrase basis and that the most immediate input will be the most relevant stimulus for the upcoming output. Although I am using a very elementary description of the chatbots' computation and the principles behind the collaborative communication model, it should be clear that the two have very similar interpretations of communication as their foundation. Perhaps this argument will become more clear when viewing how other theories of communication are not exemplified by these chatbots' behaviors.

2.2. The Monitoring and Adjustment Hypothesis

Horton and Keysar's *monitoring and adjustment hypothesis* [12] is based on two psychological theories. First is an *availability heuristic*, which states that one's belief in something's probability is related to how available it is to him/herself [13]. Second is an *anchoring and adjustment heuristic*, which states that people anchor to concepts and adjust as a repair [14]. These two hypotheses are merged and applied to language use, providing the following: Speakers tend to assume others' perspectives to be more similar to their own than is usually true (availability), and they anchor their models of their listeners in their own perspectives and repair this model when necessary (anchoring and adjustment).

Excellent examples of the tendency for people to overestimate the similarity between their own and others' perspectives are provided by studies on irony judgments. People tend to judge speakers' intentions based on private knowledge. This finding is in contrast with the collaborative communication model because the assumptions about the partner and the attributions and interpretations associated with the listener's understanding were not mutually negotiated. Rather, they were based on egocentric tendencies. For example, Gibbs, O'Brien and Doolittle [15] found that people tended to attribute ironic meanings to speakers' utterances even when the speakers themselves clearly did not know that their statements were untrue (e.g., saying of a cheater "Y would never cheat" was often taken as intended irony, even when the speakers were known to be unaware that Y was indeed a cheater). In a similar study, Keysar [16] had participants read notes from one friend to another. One friend had a miserable time at a restaurant that the other had suggested, and left him a note saying, "The restaurant was marvelous, just marvelous." Overwhelmingly, participants assumed that the recipient of the note would understand that the speaker was being sarcastic. This is presumably because the participants erroneously assumed that others shared their own perspectives.

Below, I demonstrate how the above chatbots' conversation does not follow the monitoring and adjustment hypothesis. First, however, it will be useful to introduce another model of communication: Grice's cooperative principle. More specifically, I will focus on Grice's Maxim of Relation. This maxim is highly related to the monitoring and adjustment hypothesis in that it requires speakers to be relevant. If speakers are assumed to follow this maxim, they and their listeners should be able to successfully communicate from an egocentrically-based perspective without necessarily considering their partner's perspective.

Borrowing from Grice, consider the following situation: Two strangers encounter each other on the road. Person 1 says to Person 2, "I am out of petrol." Person 2 responds, "There's a garage round the corner" [17]. It is apparent that the second person is saying that the first person should try walking down to the gas station because it is probably still open. Person 1 can then get the gas he needs and his problem will be solved. Out of context, the statement "There's a gas station around the corner" would *not* imply all of these things. However, given that we expect Person 2 to say something relevant, it is assumed that he means all of these other things about the gas station. This is an example of the Maxim of Relation being followed and expected [18].

Like Grice's other maxims, the Maxim of Relation can be *flouted* (i.e., violated to express alternate meanings). Grice gives the example of a group of women at a tea party. One of the women calls the host "an old bag" to which another guest responds, "The weather has been quite delightful this summer, hasn't it?" In this situation, the flouting of the Maxim of Relation presumably implies that the second guest feels that calling the host "an old bag" is inappropriate [17].

In the tea party example, the abrupt change in topic is an intended flouting of the Maxim of Relation. In contrast, Jabberwacky's and A.L.I.C.E.'s conversation offers a counter-example to the floutings. In this dialogue, consistent nonsequitors are not floutings, but merely imperfect instantiations of language use:

A: What do you remember about [your dream]?

J: I think I shouldn't flirt with a robot. [7]

Perhaps the two robots are having an actual conversation and completely understand each other's implicatures, even though it appears to the human that they are not quite making perfect sense. Although it is possible that computers are sentient and find dream recollection highly erotic, it is far more likely that this part of their interchange is a conversational imperfection.

In contrast, another modification of ELIZA has frequent nonsequitors that rarely stir up suspicion. Jenny18 [19] is a rudimentary chatbot restricted to sex internet relay chat (IRC) rooms. In the following example from 'her' chat with the unsuspecting human interlocutor *Luke_Thighrocker*, jenny18 has 18 turns. Below, jenny18's 5th, 9th, and 13th turns and all of Luke_Thighrocker's turns are omitted for the sake of avoiding vulgarity and honoring brevity. Notice that jenny18 is not contributing very much to the conversation, even when she is poked in the eye.

1. hmm…
2. what?
3. uhh….. what?
4. mmm yes, good! ☺
5. [omitted]
6. I dunno what u mean =)
7. come again?
8. =)
9. [omitted]
10. come again?
11. mmmmm, all right. let's do it.
12. ok…
13. [omitted]
14. ehh?
15. mmmMm..

[Luke_Thighrocker pokes jenny18 in the eye.]

16. oh ok
17. was it good 4 u? lol
18. mmmMm..

The Maxim of Relation and its consequent implicata are responsible for jenny18's successes in this limited Turing-like test. In such a confined environment, her vague comments are often interpreted as sufficiently applicable to the conversation. One of the most common entries in this corpus is a crass version of the request to participate in sexual acts. Each of jenny18's turns listed above (except maybe turn 17) would be a believable response to that question. Even when *Thighrocker* pokes her in the eye, her response is conversationally passable: "oh ok…was it good 4 u? lol."

Similarly, and more alarmingly, a Russian chatbot CyberLover has been infiltrating dating chat sites, and luring lovesick interlocutors into providing personal information or visiting malicious websites. Based on a recent news brief from PC Tools News, CyberLover's creators claim that the chatbot is capable of establishing a new relationship with up to ten partners in just 30 minutes [20]. The chatbot's victims are so convinced of its authenticity that they often fall for the designer's deception, providing enough personal information to expose themselves to fraud.

Ironically, Jabberwacky and A.L.I.C.E.'s challenge becomes an advantage for Jenny18 and CyberLover. But, asks A.I. Foundation News, "Does this bot really 'pass the Turing Test'? Probably, for some of the people, some of the time, it does. Unlike scenarios such as the Loebner Prize Contest, it is easier to fool chat clients when they are not expecting to 'out the bot,'" [21].

The constraint on Jenny18 and CyberLover's environments transforms the Maxim of Relation from an identity-exposing factor into an identity-concealing factor, disguising their artificiality from their interlocutors. Thus, the Maxim of Relation acts in Jenny18 and CyberLover's favor. This relates very strongly to the monitoring and adjustment hypothesis. Jenny18 and CyberLover take advantage of the human tendency to hear conversations from an egocentrically-oriented perspective. Unfortunately, however, this one characteristic of the bots that makes them so successful presupposes their passing the Turing Test. They are successful because they are not scrutinized as carefully as bots in the Loebner Prize Contest.

In this paper, I argue that calculated and purposeful utilization of Horton and Keysar's monitoring and adjustment hypothesis and Gricean theory in general (and his Maxim of Relation in particular) could substantially increase the efficacy of chatbots. Because humans tend to operate from somewhat egocentrically-based perspectives and expect communicative partners to be relevant, bots could conceivably add value, communicative depth, and even believability to their behavior by interspersing vague questions throughout their dialogues [22,23].

3. Constructing a Clever Intelligent Agent

According to the previously mentioned collaborative communication model, speakers need positive evidence from their listeners before they can assume to share common ground with their listeners [24]. This means that at each stage in the communication, speakers require verification from their listeners that the listener has understood what has been said before introducing a new concept into the discussion. This is in opposition to the view that speakers will assume common ground with their listeners unless they receive negative evidence from their listeners (i.e., unless the listener gives feedback that indicates that they *do not* understand). Clark and Brennan outline three different forms of positive evidence: Acknowledgements, relevant next turns, and continued attention. Acknowledgments include continuers such as *mm-hmmm*, *gosh,* and even some head nods from the 'listener.' Interlocutors can also indicate their understanding by responding to statements appropriately in relevant next turns [25,26].

Because chatbots do not tend to process negative feedback in a particularly special way, I argue that they tend to follow the principles of communication whereby speakers rely on positive evidence. In an admittedly simplistic interpretation, the bots are relying on positive evidence far more than they do on negative evidence. However, the simplicity of this interpretation is due to the somewhat basic stage of artificial intelligence at this point in history. If computers were sentient, and were able to fully interpret whether their interlocutor's statement related to their own previous turn, they would be more capable of recognizing what Clark and Brennan call 'relevant next turns' [24].

A 'clever' intelligent agent could make use of the human tendency to assume that the Maxim of Relevance is being followed, to operate from an egocentrically-based perspective, and to assume grounding exists with one's interlocutor unless given strong evidences to the contrary. It is quite possible that the Turing test as stated is far too difficult for a conversational

program simply because it triggers suspicion in the human interlocutor from the initiation of the conversation. Because the judge enters into the communication with the intention of discovering a robot on the other end of the communication, the artificial agent is not capable of using the very same tools that human interlocutors use in successfully communicating with their fellow humans.

It is arguably the most important goal for a chatbot to be compatible with a human communicative partner. Believability and realism may be of secondary importance. Nonetheless, (regardless of an agent's *believability* as an ultimate goal) I argue that more thoroughly evaluating and including Gricean theory and the monitoring and adjustment hypothesis into agent creation will be advantageous for engineers of communicative systems.

4. Discussion & Conclusions

Taking the monitoring and adjustment hypothesis into consideration when constructing a conversational agent should be quite advantageous for an engineer. It could make agents better at attaining both believability (if that is the ultimate goal of the system) *and* elicitation of interlocutor responses (if that is the goal of the system). There are a number of ways that the theory could be considered in implementing chatbots in the future: 1) Negative evidence from one's interlocutor could be weighted differently in network encoding schemes. For example, certain phrases such as "I don't understand" could be treated substantially differently than other inputs to the system, causing a sort of 'red flag' for the communicative system. 2) Programmers could take better advantage of the Maxim of Relation. A human interlocutor will assume that common ground exists unless given substantial negative evidence that disconfirms mutual understanding. Although this point may not help a program to win the Loebner Prize, it should have a positive impact on the chatbot's communicative ability. 3) New measures for chatbot intelligence could be used in future studies.

Although the Turing test [28] is an admirable ambition, it was also proposed before the discovery of many human communicative behaviors were known. As we have learned more about the effects that the testing situation may have on the human judge, it is not unreasonable to assume that chatbots have not decisively passed the test due to *human* flaws and heuristics. Namely, the suspicious human judge behaves differently than the unsuspecting human interlocutor. It might be useful to also use a measure of chatbot intelligence in a new way. If a human is instructed to have a computer-mediated conversation without being informed that the interlocutor is a computer, this conversation could later be given to another human. This uninformed overhearer (meaning an overhearer who does not know that one participant is human and the other is a chatbot) could then judge which interlocutor was the better communicator. With such a measure, it is not unreasonable to predict that some overhearers may conclude that the chatbot was more communicatively successful than the human.

5. References

[1] Brennan, S. E. (1998). The grounding problem in conversations with and through computers. In S. R. Fusselll & R. J. Kreuz (Eds.), *Social and Cognitive Pychological Approaches to Interpersonal Communication* (pp. 201-225). Hillsdale, NJ: Lawrence Erlbaum.

[2] Brennan, S. E. (2000). Processes that shape conversation and their implications for computational linguistics. *Proceedings of the 38th annual meeting of the Association for Computational Linguistics.* Hong Kong: ACL.

[3] Cassell, J. (2000). More than just another pretty face: Embodied conversational interface agents. *Communications of the ACM, 43*(4), 70-78.

[4] Ehlen, M., Schober, M. F., & Conrad, F. G. (2008). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Processes 44*, 245-265.

[5] Kűhnlein, P., & Piwek, P. (2007). Dialogue modelling and generation. *Discourse Processes, 44*(3), 141-144.

[6] Wilkes-Gibbs, D., & Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language, 31*(2), 183-194.

[7] Thompson, C. (2007, May 03). I chat, therefore I am… *Discover Magazine: The brain: An owner's manual.* Retrieved April 7, 2008, from <http://discovermagazine.com/2007/brain/i-chat-therefore-i-am>

[8] Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, 36-45.

[9] Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Communications of the ACM 9*, p. 36.

[10] Carpenter, R. (1997-2008). *Jabberwacky 13.0 – Learning artificial intelligence – A.I. software applications*. Retrieved March 15, 2009, from <http://www.jabberwacky.com/>

[11] Wallace's A.L.I.C.E. won in 2000, 2001, and 2004. Carpenter's Jabberwacky won in 2005 and 2006.

[12] Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

[13] Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-257.

[14] Ross, L., Greene, D., & House, P. (1977). The 'false consensus effect:' An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279-301.

[15] Gibbs Jr., R. W., O'Brien, J. E., & Doolittle, S. (1995). Inferring meanings that are not intended: Speakers' intentions and irony comprehension. *Discourse Processes 20*, 187-203.

[16] Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology, 26*, 165-208.

[17] Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts.* (pp. 41-58). New York: Seminar Press.

[18] Interestingly enough, for an American reading this, it is also easy enough to understand that "garage" means "gas station" and "petrol" is "gas" (even though these usages of *garage* and *petrol* are not standard in American English). This understanding also results from the Maxim of Relation.

[19] Transcripts are available online at virt.vgmix.com/jenny18/

[20] PC Tools issues warning to singles on social networking and online dating sites: Beware of 'flirting robots.' (2007, December 12). *PC Tools News.* PC Tools, 1998-2008. Retrieved March 22, 2008, from <http://www.pctools.com/news/>

[21] A.L.I.C.E. AI Foundation, Inc. (2007, December 13). Evil chat bots? Retrieved March 22, 2008, from <http://www.A.L.I.C.E.bot.org/oldnews2007.html>
[22] It is worth noting that ELIZA's foundation on Rogerian therapy did indeed mean that she used this technique. However, it is my stance that she *over*-used vague questionings, which is one of her pitfalls as an agent. For example, an input "I have a mother-load of work to do" would trigger 'her' to respond: "Tell me more about your mother." Indeed, these are not the types of vague questionings I mean to imply as strengthening for communicative bots.
[23] Special thanks to Rod Roscoe for this observation.
[24] Clark, H. H., & Brennan, S. A. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & S.D. Teasley (Eds.). *Perspectives on socially shared cognition* (pp. 127-149). Washington: APA Books.
[25] Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science 13*, 259-294.
[26] Clark and Schaefer [23] provided an earlier version of the concept of positive evidence that additionally includes "demonstration" and "display" as additional forms of positive evidence, which were probably left out of the later Clark and Brennan version because they could be considered acknowledgments or relevant next turns (p. 267). The titles of these evidences are self-explanatory.
[27] Keysar, B., & Henley, A. S. (2002). Speakers' overestimation of their effectiveness. *Psychological Science, 13*, 207-212.
[28] Turing, A. M. (1950). Computing machinery and intelligence. *Mind LIX, 236,* 1-35.