The Perception of Artificial Intelligence as "Human" by Computer Users

Jurek Kirakowski, Patrick O'Donnell, and Anthony Yiu

Human Factors Research Group, UCC, Cork Enterprise Centre
Cork, Ireland
jzk@ucc.ie, p.odonnell@student.ucc.ie, yiucw@mac.com

Abstract. This paper deals with the topic of 'humanness' in intelligent agents. Chatbot agents (e.g. Eliza, Encarta) had been criticized on their ability to communicate in human like conversation. In this study, a CIT approach was used for analyzing the human and non-human parts of Eliza's conversation. The result showed that Eliza could act like a human as if it could greet, maintain a theme, apply damage control, react appropriately to cue, offer a cue, use appropriate language style and have a personality. It was non human insofar as it used formal or unusual treatment of language, failed to respond to a specific question, failed to respond to a general question or implicit cue, evidenced time delays and phrases delivered at inappropriate times.

Keywords: chatbot, connectionist network, Eliza, Critical Incident Technique, humanness.

1 Introduction

There is a potentially diverse range of applications for intelligent agents in software use. Such agents have a part to play in domains such as user interfaces, the negotiation of information retrieval and organization, and electronic purchasing. These agents represent the active use of weak Artificial Intelligence units in computer use in general. The search engine "Encarta", devised by Microsoft, is a good example which demonstrates some of the applications associated with these kinds of agents. The software uses Natural Language Processing (NLP) in order to give the program ability to carry out sophisticated information retrieval tasks. This stems from the capacity of sophisticated NLP programs to parse the typed requests of human users and approach an "understanding" of these requests in terms of keywords which can then be used to search the web for relevant information and parse the resulting text in order to construct a response to a request.

"Encarta" offers an attractive user interface because it can parse, analyse and produce natural language utterances. This allows the program to produce relevant responses to human utterances in natural language form. It is this aspect of NLP agents that is the concern of this paper. In theory, a powerful enough NLP agent could offer a hugely attractive form of Human Computer Interaction in that computer use could be mediated by an agent that behaves as though it can understand instructions that are typed in by human users and respond in kind with natural-seeming utterances.

J. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2007, LNCS 4552, pp. 376–384, 2007.

The point has been raised that in order to communicate with computers, humans must learn the language of computers and that computers are, at present, incapable of communicating by using human languages [8]. The prospect of an NLP agent that is powerful enough to deal with human languages would radically change this state of affairs.

There are several ways of approaching the problem of building a software engine that has to deal with human language utterances. Since an agent that is successful in this task is essentially a form of Artificial Intelligence, it seems fitting to begin by describing the traditional approach to AI and how it relates to this problem. Traditional AI approaches rely on a Strong Physical Symbols System approach (SPSS), whereby a series of symbols is given to the engine in question, the symbols are manipulated in some logical manner within the engine and a series of symbols is given as output [7]. There are several problems attendant on these approaches, both in general and in the specific case of language processing. A general criticism of SPSS approaches is that of symbol grounding [6]. The problem is essentially that of how a system that depends on the manipulation of virtual symbols can ever ascribe any kind of meaning to those symbols, except in terms of other arbitrary symbols thus defined themselves. Another problem that is of particular interest with regards to the question of language is the problem of emergent problem spaces as associated with these types of systems. This consists of attempts by the engine to generate all possible solutions to a problem once posed, and an inability to choose the most likely solution from amongst the other contenders. This problem is especially relevant to language processing as it is a task that must occur in real time, and any attempt by a language engine to test each individual solution before choosing the correct one will be time consuming. The classic example of this type of problem in language processing comes from attempts to build sentence parsers along this line. The engine, when tested, generated five possible meanings for the sentence "Time flies like an arrow" [1].

The second prominent approach in contemporary Artificial Intelligence is the connectionist approach (also referred to as parallel distributed processing or the neural net approach). There are at least three unique features that make a connectionist network a powerful system for handling human conversation: namely superpositioning, intrinsic context sensitivity, and strong representational change. [3] Firstly, two representations are said to be fully superposed when the resources used to represent item 1 are co-extensive with those used to represent item 2. The natural mechanism of connectionist learning and superpositioning storage yield a system that will extract the statistical central tendency of an exemplar. This is usefully seen as embodying prototype-style knowledge representation. The network extracts various feature complexes and thus comes to encode information not just about specific exemplars but also about the stereotypical features set displayed in the training data. The network can generalise novel case sensibly by dint of its past training.

Secondly, the connectionist network concept can also display intrinsic context sensitivity. The most radical description of this would be that connectionist system does not involve computations defined over symbols. Instead, any accurate picture of the system's processing will have to be given at the numerical level of unites, weights and activation-evolution equations, while the symbolic-manipulating computational description will at most provide a rough guide to the main trends in the global behaviour of the system [3]. The network can then learn to treat several inputs, which result in subtly different representational states, as prompting outputs which have much in common.

Thirdly, a connectionist network can show strong representational change. Fodor [4] suggested concept learning can only consist in the triggering of innate representational atoms or the deployment of such atoms in a "generate and test" learning style. According to [3] this is weak representational change as the product necessarily falls within the expressive scope of the original representational base. The connectionist network on the other hand, can acquire knowledge without the benefit of any such resource. For example the NETtalk [10] and the past-tense learning network [9] both begin with a set of random connection weights and learn about a domain "from scratch".

Connectionist models however have a similar grounding problem as the SPSS approach does. Connectionist models explain symbols by a series of context-sensitive connections. The process itself does not 'bottom-out' or come to a definition that is *not* prone to context infection. In additional, there is a problem in systematicity [5] in that a connectionist network can fail to process sentences with constituents in novel syntactic position and at a novel level of embedding when processing includes determining the word's semantic role.

The symbol grounding problem is the problem of representing meaning in a system of purely arbitrary symbols. One approach to robotics and AI in general that may be able to address this problem involves dodging the question of representation altogether. Wallis [11] discusses the possibility of producing agents that can exhibit all the characteristics of an intelligent agent (intention, planning etc.) without using any more representation "than a microwave oven would." Wallis' stance is informed by Brooks' [2] approach to robotics, wherein robots can be developed that behave in an autonomous, intelligent fashion without any bona fide "understanding" of their own behaviours or why they are performing them. The essential tenet that underlies this approach is that an agent's intelligent behaviour arises out of an interaction between the agent and its environment, in the service of achieving some goal. Reflective reasoning about the environment, the goal or the behaviour by the agent is not necessary for the behaviour to be described as intelligent. The agents that exhibit this type of architecture can perform behaviours that can be described as intelligent without possessing any capacities that we would describe as "intelligence" because their actions make sense within their environment with regards to the satisfaction of some goal. Thus, if it is possible to produce a chatbot that has no representation of meaning but can behave as though it did (i.e. seem to understand utterances and interact with users), then users would be forced to conclude that its conduct within a dialogue was "intelligent". The earliest types of chatbot programs, that scan for keywords and match responses, can be seen as non-representational chatbot agents. It is intended to examine the way in which one of these agents will interact with users, and how it might be possible to improve on its ability to be regarded as an agent that produces intelligent language behaviour. What is interesting is whether a non-representational approach such as this can be brought to bear in an arena such as language, a system of representational symbols.

It is not appropriate in this paper to attempt to select between these three approaches to the architecture of a natural language machine. No doubt in the end the "best" approach will be a hybrid of some kind and there are problems of principle as well as of implementation. We note that in the past research has taken a particular technology as a given and focussed on the application. We propose to turn the problem round.

That is, it is intended to do a much more "requirements" orientated survey, to identify what aspects of speech comprehension and production by software agents characterise them as being "inhuman" in the eyes of computer users and which aspects are characteristic of human language behaviour. It may later be appropriate to discuss which types of programs and architectures are best equipped to support the type of behaviours seen as quintessentially "human". In other words, the research question addressed in this paper is: when users interact with an agent that is equipped to process human language and respond with utterances of its own, what kind of mistakes can the program make that a human wouldn't and that make the dialogue between user and agent seem unnatural.

2 Method

In the experiment fourteen college-aged participants were asked to interact with an Eliza-style computer program (chatbot) for three minutes and then to participate in the elicitation of critical incidents with a transcript of their session. The program was based on the classic Eliza design with two important differences. Firstly, there was no mechanism that retained previous phrases entered by the user which could be used to re-start a stalled conversation (eg: "Tell more more about [a previous utterance]".) This was for theoretical reasons as will be discussed later. Secondly, there was a mechanism which enabled the chatbot to switch contexts on detecting particular words. Thus if the chatbot detected the word "music" the whole list of trigger phrases and responses changed to a music-orientated set.

A qualitative approach incorporating the Critical Incidents Technique (CIT) and content analysis of responses is involved in this study. At the end of this interaction, the participants were presented with a printed transcript of the dialogue and asked to highlight instances of the conversation that seemed particularly unnatural (up to three examples) and then to report why this was so. The same was done for up to three examples of speech that did seem convincing. The data produced by the critical incident technique were content analysed, with user responses being sorted by theme. The data coding was cross-checked independently by another individual. Inter-rater reliability of approx. 0.53 was obtained in the first pass. Items on which there was disagreement were discussed and placed in mutually agreeable categories with the assistance of a third independent rater. We are reasonably sure that the categories that have emerged represent reproducible aspects of the data set.

3 Results

The various themes that were produced during the content analysis were as follows. Firstly, under the heading of unconvincing characteristics-

- *Fails to maintain a theme once initiated.* In that, once a theme emerged in the dialogue, the chatbot failed to produce statements relevant to that theme in the following section of the dialogue.
- *Formal or unusual treatment of language.* Some statements in the chatbots database seemed overly stiff and formal or used unusual words and language.

- *Failure to respond to a specific question.* Users would ask for a specific piece of information, such as asking the chatbot what its favourite film might be, and receiving no answer.
- *Fails to respond to a general question or implicit cue.* Users offer the chatbot a cue (in the form of a general question, like "How are you?", or offer a cue in the form of a statement, like "Tell me about yourself." Or "Let's talk about films then.") and receive an irrelevant response.
- *Time delay.* A fairly cosmetic fault, users felt that the chatbot responded too quickly to a detailed question or too slowly to a courtesy.
- *Phrases delivered at inappropriate times, with no reference to preceding dialogue.* Where generic type phrases did not fit into the conversation in a natural way, or the chatbot responds to an inappropriate key phrase, with a resulting nonsequitur.

Under the heading of convincing aspects of the conversation-

- *Greetings.* Several participants identified the greeting as a human-seeming characteristic.
- *Maintains a theme.* When the chatbot introduced a theme and was successful at producing a few statements that were relevant to that theme, users found this convincing.
- *Damage control.* When the chatbot produced a breakdown in communication (for any of the reasons mentioned earlier) and then produced a statement that seemed to apologise for the breakdown or seemed to redirect the conversation in a more fruitful direction, users found this a convincingly human trait.
- *Reacts appropriately to cue.* Users found it convincing when the chatbot responded appropriately to a cue such as "How are you?" or "Tell me about yourself."
- *Offers a cue.* Users found it convincing when the chatbot offered a cue for further discussion, such as "What do you want to talk about?" or offered a range of topics for discussion.
- *Language style.* Users found conversational or colloquial English to be convincing.
- *Personality.* The fact that the chatbot was given a name (in fact, even users who did not report the inclusion of a name as convincing referred to it as a "Sam" or a "he") suggests that users wish to assign a personal agency to the chatbot even in the teeth of discrepant knowledge.

4 Discussion

This research focuses on requirement and not any kind of implementation. For now it is enough to identify what traits in the bot-human interaction make it different to human-human interaction and how best these shortcomings might be addressed. Indeed, a reassuring symmetry emerges in the themes identified by users as being convincing or not: maintaining a theme is convincing, while failure to do so is unconvincing, formal or unusual language is unconvincing while colloquial or conversational English is the opposite. Reacting appropriately to a cue is human while failing to a react to one isn't. Delivering an unexpected phrase at an inappropriate time does not impress, but damage control statements can rectify the situation. It is time to address each feature of the bot-human dialogue in a little more detail.

4.1 Maintenance of Themes

One of the factors, upon which the success or failure of the program to appear human seems to depend, is its ability (or lack thereof) to maintain a conversational theme once introduced. The Eliza-style chatbot used in this trial has no memory of a conversation as such (it operates on a first order Markov process, whereby each token is generated in response to the token immediately preceding, with no reference to the accumulated tokens, in this case token = utterance and accumulated tokens = the whole dialogue). This does not preclude it from maintaining a theme however; indeed several participants reported its ability to do so as a convincing feature of its dialogue. The means by which this is accomplished (given that the program has no "memory" of the conversation) is now described.

The chatbot used was unlike the classic Eliza program in that as well as having specific phrases activated by the presence of a keyword, the program could activate a whole database of phrases in response to a key phrase that are specifically related to that phrase (for example, an inventory of keyword-response pairs that are related to music can be prompted by the word "music").

Thus, the program has access to a database of phrases that are most likely to be relevant to the theme raised. At present, failure to maintain a theme that has activated one of these databases may be due to the fact that these databases contain all the same generic response phrases and keyword-response pairs as the general text database that serves as the default set of responses. This makes the likelihood that a theme-relevant phrase is activated lower than if the specialised databases were to contain theme-relevant phrases only. Thus, a means of improving the ability of this program to maintain a theme in conversation might be to enlarge the number of theme relevant keyword-response pairs in these databases and remove most of the generic keyword-response pairs from these "themed" databases.

4.2 Failure to Respond to a Specific Question

This problem, essentially, is a question of how much information is contained in the program's memory and whether or not it can be accessed. Thus, if a person were to ask the program "What is the capital of France?" and the program did not have the information required, the program seems less human. There is no easy way to solve this problem. The solutions are to give the program a large enough database of information to be able to cope with most information requests of this kind (this approach suffers from the fact that the database is still a finite resource and almost certainly contains less information than a human would be expected to) or to grant the program access to the internet and equip it with a more powerful means of parsing information requests so that it can then establish the exact nature of a request and search for the relevant data on the internet. This first solution is brute force and is probably most relevant to a personal-use "humanised" AI, with a role as a user-interface for small-scale personal computer use, while the second is the type of approach that might be associated with a general information retrieval agent such as Microsoft's "Encarta".

4.3 Responding to Social Cues

This category covers the failure or success of the program to react appropriately to a social cue such as "How are you?" or "Tell me about yourself." Some of these cues can be treated in a similar way to the information requests dealt with above, in that an appropriate response can be matched, from a database, to a specific cue.

4.4 Formal and Colloquial Language

In general, formal language was regarded as being an unconvincing trait of the program's, with casual or colloquial language being preferred. Replacing formal phrasings with casual equivalents is a relatively minor adjustment that can be made to improve the program's performance. It is worth bearing in mind however that this trial involved a chatbot that was geared towards free conversation as opposed to being a helper agent in a structured task. In other circumstances, language style might not be a consideration for users at all, or perhaps even more formal and precise language might be preferred (eg in making a financial transaction.)

4.5 Greetings and Personality

Some users reported that certain surface details involved in the chatbot's dialogue made it seem more human by their very presence. The fact that the bot "introduced itself" at the beginning of the dialogue and was given a human name for the trial influenced people into regarding it as slightly more human. This is separate from the functional issues involved in recognizing conversational breakdown and issuing damage repair, and is probably more related to personal preferences.

4.6 Offers a Cue

The chatbot was deemed to be very "humanlike" when it offered cues on which users could elaborate. The possibility has already been raised of including more cues which are designed to elicit clarification in situations where the chatbot does not have enough information to respond appropriately to a cue. This promotes information exchange between the user and the chatbot and is likely to reduce ambiguity and allow the chatbot to react more reliably to user-statements.

4.7 Phrases Delivered at Inappropriate Times

This is an enduring problem of the Eliza style keyword-response chatbot, generic phrases are produced which do not fit well into the conversation, or a keyword prompts a response that is inappropriate in the context it is used. The first problem can be caused when the generic "placeholder" phrase is a poor one. In the case of the second problem, the chatbot might produce an inappropriate phrase due to the fact that it is insensitive to context. A word which means one thing in a certain context, and which prompts an appropriate response, might mean something completely different in a different context and the same response, when prompted, will no longer be appropriate. Some suggestions for remedying this problem are to equip the program with statements that ask for clarification and to refine the types of keywords that prompt particular responses. In addition, a chatbot that relies on a connectionist architecture may well be more sensitive to context than the model described here and may thus be able to select appropriate responses with a high degree of accuracy.

4.8 Damage Control

In certain situations, the chatbot seemed to be offering to change the topic of conversation after a particular line of conversation broke down, or to try and clarify previous statements. This is a further example of the kind of information-exchange that can occur between users and agents. Not only does this ability seem to make the chatbot appear more human, it would be a valuable ability to develop in any of the major potential applications of chatbots as helpful agents. This type of capability would allow for a more refined search when using information-retrieval agents. In personal computer user-interfaces, this kind of information-exchange opens up the possibility for the agent to make suggestions as regards computer-use.

With regards to the method of analysis employed in this study, it is intended to discuss the level to which the Critical Incident Technique was an appropriate tool of assessment in this trial. The benefits of the Critical Incident Technique as regards this study were as follows:

- Rare events were noted as well as common events, thus the situation in which bot-human interaction could break down and then be retrieved by the bot in a damage control exercise did not occur in all or most of the dialogues but it was identified alongside more common shortcomings of the bot nonetheless.
- Users were asked to focus on specific instances of communication breakdown (as opposed to being allowed to offer the vague opinion that the dialogue "felt wrong") and this allows for a more precise focus on individual problem areas (such as being able to treat "failure to answer a specific question" as a separate problem to "failure to respond to a general question or cue").

However, some shortcomings of the Critical Incident Technique as used in this trial were as follows:

- There is no indication as to the relative severity of failures by the bot to appear human. In other words, it is difficult to tell if users found the agent's inability to maintain a conversational theme a more serious problem than the delivery of unexpected and inappropriate phrases during the dialogue, or even if there is a degree of individual difference involved in which characteristics of the bot's conversation-style are pertinent to its seeming human.
- This method of analysis requires a focus on specific incidents of success or failure and is not particularly sensitive to context. This trial involved a simulated conversation, in which context would be important in establishing whether or not the dialogue seemed natural and though participants are asked to describe the events that lead up to a critical incident as part of their report, some information regarding the context of the conversation as a whole is probably missed.

References

1. Bobrow, D.: Syntactic Analysis of English by Computer – A Survey, tech report 1055, BBN (1963)
2. Brooks, R.A.: Intelligence without representation. Artificial Intelligence 47, 139–159 (1991)
3. Clark, A.: Associative Engine: Connectionism, Concepts, and Representational change. Bradford Book; London, England (1993)
4. Fodor, J.: Representations: Philosophical Essays on the Foundations of Cognitive Science. MIT Press, Cambridge (1981)
5. Hadley, R.F: Systematicity in connectionist language learning. Mind. and Language 9, 247–272 (1994)
6. Harnad, S.: The Symbol Grounding Problem. Physica D 42, 335–346 (1990)
7. Newell, A., Simon, H.A.: Computer science as empirical inquiry: Symbols and search. Commun. Assoc. Comput. Machinery 19, 111–126 (1976)
8. Pinker, S.: The Language Instinct. Penguin. London, p. 193 (1994)
9. Rumelhart, D., McClelland, J.: On learning the past tense of English verbs' in Parallel Distributed Processing: Exploration in the Microstructure of Cognition, vol. 2. MIT Press, Cambridge (1986)
10. Sejnowski, T., Rosenberg, C.: NETtalk: A Parallel Network That Learns to Read Aloud. Technical report JHU/ECC-86/01, John Hopkins University (1986)
11. Wallis, P.: Intention without representation, Philosophical Psychologyy, vol. 17(2) (2004)