'Realness' in Chatbots: Establishing Quantifiable Criteria

Kellie Morrissey and Jurek Kirakowski

School of Applied Psychology, University College Cork, Ireland
{k.morrissey,jzk}@ucc.ie

Abstract. The aim of this research is to generate measurable evaluation criteria acceptable to chatbot users. Results of two studies are summarised. In the first, fourteen participants were asked to do a critical incident analysis of their transcriptions with an ELIZA-type chatbot. Results were content analysed, and yielded seven overall themes. In the second, these themes were made into statements of an attitude-like nature, and 20 participants chatted with five winning entrants in the 2011 Chatterbox Challenge and five which failed to place. Latent variable analysis reduced the themes to four, resulting in four subscales with strong reliability which discriminated well between the two categories of chatbots. Content analysis of freeform comments led to a proposal of four dimensions along which people judge the naturalness of a conversation with chatbots.

Keywords: Chatbot, user-agent, intelligent assistant, naturalness, convincing, usability, evaluation, quantitative, questionnaire, Turing, Chatterbox.

1 Evaluating for Naturalness

Conversational agents, or chatbots, are systems that are capable of performing actions on behalf of computer users; in essence, reducing the cognitive workload on users engaging with computer systems. There are two key strategies used. The first is the use of a set of well-learnt communicative conventions: natural language and the accepted conventional structure of a conversation so that the user does not need to learn artificial conventions (such as SQL, other query languages, or highly constrained programming methods.) The second is enabling the user and the computer to refer to broad shared classes of knowledge of which either the computer, the user, or both hold the specific details so that the solution to a problem can be arrived at by negotiating through the knowledge space in a way that neither side need concern themselves with details which are difficult or impossible for that side to represent.

Implementation of these two strategies: naturalness of interaction and sharing knowledge space are the two essential features of all conversational agents. Different agents vary in the success of their implementations of each. But the important point is that as far as the user is concerned, the interface is one: an intelligent conversation heeds conventional structure as well as being about a shared referent.

Chatbots are over fifty years old. Turing [19] in his famous thought experiment set up what is considered to be the touchstone of evaluation, despite some researchers'

M. Kurosu (Ed.): Human-Computer Interaction, Part IV, HCII 2013, LNCS 8007, pp. 87–96, 2013.
© Springer-Verlag Berlin Heidelberg 2013

claims (e.g., [10],[11]) that Turing's test offers neither an operational definition of intelligence in computers nor necessary and sufficient conditions for the demonstration of such intelligence. Weizenbaum in 1966 [20] wrote his ELIZA in the MAD-SLIP programming language and demonstrated how one could simulate human conversation through simple pattern-matching of user input to the data stored in its script. Weizenbaum's complaint that ELIZA was often not seen as the trick it really was and his vision that a more successful ELIZA would build a model of the person with whom it converses during the conversation is perhaps a symptom of the representation-dominated theories of the mind current at the time, embodied most starkly in Fodor's landmark book *The Language of Thought* [9].

There is a potentially wide range of applications for chatbots today. These types of agents have a part to play in domains involving the negotiation of information retrieval and organization. As the amount of knowledge held in data stores expands, and as the technical level of skill required by the average user diminishes to make this knowledge available to an increasingly diverse user population, intelligent agents become increasingly important to the universal acceptance of technology. Chatbots are found in stores and help sites as embedded online assistants, in chatrooms as spam agents, and in video games as non-playing characters. Notable applications of recent chatbot-like technology in the press are IBM's Jeopardy-winning computer Watson [5] and Apple's embedded "personal assistant", Siri [1].

So what makes for a convincing, satisfying, perhaps a natural interface for a user agent?

It is perhaps to answer this question that challenges such as the Loebner Prize still run Turing-like tests each year in an attempt to spur on the creation of a chatbot that can converse in a naturalistic fashion. The Loebner Prize was started by Dr Hugh Loebner in 1991 and has been held every year since then, with multiple entrants each year. While the prize money of $100,000 which has been set aside for the winning chatbot undoubtedly inspires many programmers to create and improve their chatbots, the Loebner Prize has been criticised for a lack of realism. Shieber [18], in attendance at the first Loebner Prize, contends that the Loebner Prize is not a true representation of Turing's test: the conditions of the challenge are modified extensively in order to allow the chatbot what is considered to be a fighting chance – the topic of the human-computer conversation itself is restricted to a singular domain and is thus not a free test. The binary "yes/no" decision of the original Turing test, Shieber observed, is also replaced by a ranking format, in which the judges rank in order of their "humanness" but not specifying an absolute "human" threshold. Is there really a point in running a test in which such large concessions need to be made?

Cohen, in a paper entitled "If not the Turing test, then what?" [7], describes a number of differential intelligence tests for bots. Cohen suggests such trials be drawn from the sort of tasks that third graders in a North American elementary school ought to be able to complete. When considering the sort of tasks that Cohen suggests, it is interesting to note that the majority of these tasks require a significant ability to understand, manipulate and produce concepts behind language. While this kind of ability is considered important to demonstrate intelligence in humans since the days of Binet's pioneering demonstrations [3] it is doubtful whether this is even relevant to the evaluation of naturalness in chatbots.

However, Cohen's *criteria for a good challenge* are fully in accord with the aims of the research presented here. They are threefold:

1. such a challenge must produce feedback that is more developed than the non-gradated "yes/no" of the Turing test,
2. it must have a sense of monotonicity, allowing for repeated reproductions of the challenge to verify the results of previous challenges, and
3. it must "capture the hearts and minds of the research community" - while the Loebner Prize and the Turing test have certainly engendered a large amount of discussion, very few working in the area of intelligent systems would seriously put Cohen's *specific tasks* forward as a measure of the results of chatbot development.

Shawar and Atwell [17] propose "glass box" and "black box" methodologies in order to assess a chatbot. These methodologies represent two sides of appraisal: glass box methodologies assess a given conversation technically for grammar, syntax, sentence structure and appropriateness of answers; while "black box" methodologies broadly attempt to measure user satisfaction. In testing these methodologies using a goal-based task and an Afrikaans-literate chatbot, Shawar and Atwell found that such a separation was ill-suited to the task at hand and proposed that the Loebner Prize criterion of naturalness is in the end perhaps preferable. Their final suggestion is that chatbot success should be functionally defined: "the best evaluation is based on whether it achieves that service or task." In general, we would agree with such a task-based criterion. But chatbots are no longer predominantly used for work-based tasks in the sense of the ISO 9241 part 11 definition of usability [13]. So what to do then, when the chatbot may be designed for nothing more than to be a partner in an amusing natural-seeming human conversation?

Semeraro et al. [16] used a top-down approach to evaluate their agent-based interface, constructing a questionnaire which assesses the chatbot's ability to learn and to aid the user, its comprehension skills, ease of navigation, effectiveness, impression and command. Hung, Elvir, Gonzalez and DeMara [12] note that this is a subjective approach: a criticism which bolsters the need for a statistically reliable evaluatory instrument. They also note that it is more of a general indicator of performance, rather than an appraisal which would lead to generalisable findings for chatbots. Rzepka, Ge & Araki [15] use a similar 1-10 rating system assessing naturalness and technical ability to continue a conversation in assessing the performance of older-style ELIZA chatbots and newer commonsense retrieval bots, which was then expressed as a "naturalness" degree and a "will of continuing conversation" degree. The issue with these methodologies, however, is that the scales and questionnaires used to test the chatbots are not themselves verified as sufficient by ordinary users and lack reliability and validity as measuring instruments.

The research question addressed in this paper is part on an ongoing research programme to generate measurable criteria for the naturalness of chatbot dialogue that are acceptable to people who are more interested in the results of chatbot development than the technical issues of the development itself. Although at this stage we feel we can make a modest contribution to our knowledge on the subject we still have a way to go, as we will explain in the conclusion to this paper. There are two studies which we wish to present.

2 Study 1: In the Words of the Users

This study used a chatbot that was modelled on the ELIZA scheme but which explicitly made use of situational semantics [2]. Information, according to this view, exists in situations - which are usually local and most probably incomplete. Users, who in this case are considered to be the environment within which the chatbot finds itself, could be considered to have a large amount of information that is part of the situation and which therefore does not need to be represented in the software databases. Conversations can be created according to topics, in which there may be a number of types, and the contents of conversations are ELIZA-type input-output transformations, which are considered as tokens linking to the types in the conversation topic. The program worked on a simple subsumption architecture [4], in that there are three layers, each of which could hold the floor at any one moment and which communicate with the other layers by very simple excitatory signals.

- Layer 1: Conversational maintenance on a given topic where tokens are connected to types and types are connected to other types;
- Layer 2: A switching agent to find a new topic and connect to it;
- Layer 3: General purpose social control: phatic conversational tokens.

The program very explicitly did not store previously unused keywords to pop back if it got stuck, the way ELIZA did: instead firstly social control sought to bring the conversation back to track, and then after a while the switching agent came in to negotiate a new topic with the user. Many tokens were created by the developers of each type, and tokens were selected by the program on a pseudo-random basis from each type when required. A randomly-generated time interval preceded each response by the chatbot. The chatbot was given the name of “Sam” with no particular acronym in mind.

In the experiment fourteen participants were asked to interact with the chatbot as described above for three minutes and then to participate in the elicitation of critical incidents with a transcript of their session. Participants were all tested individually in a HCI lab with one experimenter present. No participant was under the illusion that they were communicating with anything other than a chatbot after a few exchanges, although no explicit cues were given by the experimenters. Respondents were tested in the vicinity of a half-silvered mirror behind which a dialogue partner might have sat.

The Critical Incident Technique [8] requests the respondent to identify particular moments in an experience that the respondent, in hindsight, considers to have been critical during the experience. At the end of the interaction, therefore, 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 were asked to explain why this was so. The same was done for up to three examples of the dialogue that did seem convincing. By the time the respondents had marked the transcript up and been interviewed, it was very clear to each respondent what they had participated in.

The data produced by the critical incident technique were analysed by content. Three raters participated overall. No particular brand of qualitative analysis was considered to be specifically appropriate, although Grounded Theory [6] might come closest. The first rater went through the user responses and identified each response as belonging to one specific theme to do with having a conversation. No themes were created *a priori*, they emerged as a best fit from the data. The data coding was cross-checked independently by a second rater. Inter-rater reliability of approx. 0.53 was obtained in the first pass, which is low (but not usually assessed in Grounded Theory approaches anyway.) Items on which there was disagreement were discussed and placed in mutually agreeable categories with the moderation of a third independent rater. The researchers were reasonably sure in the end that the categories that emerged represented reproducible aspects of the data set.

In general, the kinds of comments reflected the strengths and weaknesses of the three-layer architecture of the chatbot as implemented, and also showed that respondents in general thought that both communicative conventions and the shared knowledge space were of concern when considering the naturalness of the conversation.

A reassuring symmetry emerges in the themes identified by users. For instance, being convincing or not: maintaining a theme is convincing, while failure to do so is unconvincing; colloquial or conversational English is convincing while formal or unusual language 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. This research was reported by Kirakowski, O'Donnell and Yiu in 2009 [14] who give a full account of each of the seven themes extracted. They are, in summary:

1. Maintenance of themes
2. Responding to a specific question
3. Responding to social cues
4. Using appropriate linguistic register
5. Greetings and personality
6. Giving conversational cues
7. Inappropriate utterances and damage control.

However, there is no indication as to the perceived relative severity of failures by the chatbot. In other words, it is difficult to tell if users found the chatbot's inability to maintain a conversational theme to be a more serious problem than the delivery of inappropriate utterances during the dialogue, or even if there is a degree of individual difference involved in which characteristics of the chatbot's linguistic register are pertinent to its seeming to be natural.

3 Study 2: Towards Quantification

This study developed the first draft of a questionnaire. Questionnaire statements (or items) were created following the structure of the seven themes in Kirakowski, O'Donnell & Yiu's paper ([14], henceforth called the KO'DY structure.) For each theme, at least 4 statements were initially generated that attempted to capture the substance of the theme. Two preliminary validation steps were carried out.

Firstly, face validity of the items was assessed in a meeting of a research group attended by 12 experienced researchers and postgraduate students in the field of cyberpsychology, many of whom also had psychometric expertise. Secondly, the pool of items was then reassessed according to the HCI literature by following articles published relevant to the keywords in the items, in order to increase construct validity. The final inventory consisted of 23 items, with 2-4 statements attempting to measure each KO'DY theme. Items were randomised so as to avoid order effects.

The answering format was a five point frequency scale with the anchors “always”, “often”, “sometimes”, “seldom” and “never” - the rationale for this five point scale was that it matched the statement format of the items. An open-ended question at the end of each evaluation form asked the participant “what do you think this chatbot is best at?”

Participants were chosen from the undergraduate population of University College Cork, Ireland, and were 11 male and 9 female between the ages of 19 and 30 with a mean age of 23. They were all fully briefed as to the nature of the experiment.

Chatbots were chosen on the basis of their placing in The Chatterbox Challenge; an annual chatbot competition along the lines of the Loebner Prize. Five winning entrants were chosen from the 2011 competition to act as the “good” chatbots and five entrants which failed to place in the 2011 competition acted as the “poor” chatbots. As multiple independent judges assess these bots in the Chatterbox Challenge, test-retest reliability should be good, as should, one hopes, be the case for objectivity. Each participant had to evaluate all ten chatbots. The chatbots were presented in a Latin Square design to minimise order and sequence effects.

The apparatus used was a Dell computer running Windows 7 and Google Chrome connected by fast ISDN to the Internet. A basic HTML interface was created for the purposes of the study, briefly listing instructions for the participant and containing internet links to the ten chatbots the participant was to encounter. Clicking on each link in turn opened a new tab which then loaded the page in which the chatbot was embedded. Participants chatted with each chatbot for five minutes on a topic of their choice. After five minutes they filled out an evaluation questionnaire for the chatbot and went on to the next.

Although it would have been straightforward to compute an overall score for each chatbot by summing the 23 items, computing reliabilities, and carrying out an analysis of variance, we were tempted to go slightly further and to attempt to find out how the matrix of 10 x 20 x 27 data items factorised, expecting to find a factor structure similar to the KO'DY structure. We are aware that because each respondent is each responsible for 10 questionnaires within the dataset there may be an amount of spurious intercorrelation between the chatbot scores which is impossible to estimate - but which might make the matrix more difficult to solve because of multicollinearity. We thus present these results with some reservations.

We put the data through a Principal Components Analysis (PCA, using the SPSS 18 package): PCA highlights the existence of linear components in the data set and assesses the percentage variance that these components contribute to the overall variance. This serves to narrow the scope of the statistical analysis. The contribution of eigenvalues, scree plot and item interpretation allows for an initial data reduction at this stage of statistical analysis. A varimax rotation was then utilised. This transformation of factor loadings allowed for a clarified interpretation of the results.

The initial correlation matrix contained many coefficients of .4 and above. The Kaiser-Meyer-Olkin measure indicated the sampling adequacy for the analysis, KMO = .92, and Bartlett's Test of Sphericity achieved statistical significance, 2303.853, $p < .001$, verifying that correlations between items were sufficiently large to justify PCA. The best solution which made sense was with four factors, this explained 59.86% of the variance (we tested from two up to eight factors but the scree plot of eigenvalues showed a definite bend after four and the semantics of the rotated factors supported this.) Varimax rotation on the four factors showed a clear, simple structure. We further conducted a Cronbach's alpha coefficient for reliability and found all four scales achieved 0.70 or higher, thus satisfying the lower level of reliability criteria for scales suitable for research purposes.

The factors are as follows:

1. Factor one, broadly labelled **Conscientiousness** is the largest factor, comprising of ten items which measure how the chatbot seems to keep track of the conversation at hand and how appropriate its responses were. Conscientiousness had a high Cronbach's Alpha of 0.915.
2. The second factor was labelled **Manners**, consisted of 6 items and assessed the ability of chatbots to display polite behaviour and conversational habits. Greetings, apologies, social niceties and introductions were constructs measured in the items within this factor. The Cronbach's alpha coefficient for the factor was .763.
3. The third factor was labelled **Thoroughness** and consisted of 4 items, measuring the formal grammatical and syntactical abilities of the chatbot. The Cronbach's alpha coefficient for the factor was .726. This factor had a large effect size (Cohen's d= 1.2) when we came to analyse differences between chatbots (see below), which suggests that the figure for alpha is depressed simply because there are only 4 items in the factor and that this is an important factor.
4. The fourth and final factor was **Originality**, consisting of three items which measured the chatbot's ability to produce what seemed to be original material and also its ability to take the initiative in conversations. The Cronbach's alpha coefficient for the factor was .735 and it is most probably also affected by the small item size.

Given that these four extracted factors seem to have good reliabilities, we then conducted a 2 x (5) x 4 way analysis of variance to establish whether the overall questionnaire was able to distinguish between good and poor chatbots, and whether there were any differences in the profiles of the average good and average poor chatbot. There was a significant main effect of quality of chatbot, meaning that overall the questionnaire does discriminate well (p < 0.01). There was also an interaction between quality and scales (p < 0.01) in which Thoroughness gave rise to the biggest difference: in other words, Thoroughness is the biggest discriminator between good and poor chatbots. Manners gave rise to the smallest difference, although it was still statistically significant (p < 0.01) as were the differences on all the factors.

The open-ended questions "what do you think this chatbot is best at?" were coded for content analysis; again, tending towards a Grounded Theoretic approach [6]. Codes adhered closely to the data and attempted to explain, conceptually, what was occurring in each line of the participants' answer. In all, four salient themes were present in the data.

Asking and Answering: The transactional nature of the conversation taking place between user and chatbot is highlighted here.

Originality: Users often point out the originality (or lack thereof) of some chatbot responses.

Personality: Participants refer often to the chatbot as having a "personality": the issue of manners and politeness in chatbot discourse is one which arises time and again.

Relationship with User: An interesting theme which may point towards somewhat of a sense of intersubjectivity between chatbot and user – however illusory that intersubjectivity may be!

4 Combining the Results and the Next Step

A picture of what non-technical users are expecting from a chatbot is beginning to emerge, although the final step is to revise the current 23 items by adding items suggested from the content analysis of the second study to balance up the scales, revising those items from the first study which loaded less well on the four original factors, and then conducting an exploratory followed by a separate sample confirmatory study. If all goes well, we should, by the end of the confirmatory study phase be able to offer a relatively short instrument with high reliability, validity, and a reference database against which we can score the percentile of naturalness at which a chatbot performs. At present we are not able to do this, and we reserve not to publish the current item bank and its loadings on the grounds that it is work in progress.

However, as a summary, we can see four broad dimensions on which the user might judge the naturalness of a chatbot. These may to some extent be inter-correlated so that a chatbot which does well on one will also do well on one or more of the others.

A Chatbot should be Conscientious. It should be able to keep track of the conversation, attend to the flow of the conversation, maintain themes, pick up appropriate cues, and ask and answer pertinent questions.

A Chatbot should Display Originality. It should have some interesting information about the conversational theme (*specialized* in addition to *common* topics), and it should be able to take the initiative in conversations at times, perhaps by suggestions to change to related themes.

A Chatbot should Display Manners. It should show that it has good conversational habits, can do damage control if a conversation seems to be losing its way, and should maintain an appropriate (perhaps friendly) personality and develop a relationship with the user.

A Chatbot should be Thorough. It should use appropriate grammar and spelling consistently, and consistently adopt an appropriate linguistic register with the user.

As to how such a perfect chatbot should be coded, we are quite agnostic on this point and proponents of the three major approaches to design (Strong Physical Symbols System, Connectionist, or Situational Semantics) will have their own solutions. We favour the Situational Semantics/ Subsumption architecture of Sam and note that the amount to which each of the dimensions is incorporated in each layer will vary with respect to what that layer does; but that overall, the machine should incorporate all four dimensions. This, after all, is what our users tell us they expect.

References

1. Aron, J.: How innovative is Apple's new voice assistant, Siri? New Scientist 212 (2386), 24 (2011)
2. Barwise, J., Perry, J.: Situations and Attitudes. Mass: MIT Press, Cambridge (1983)
3. Binet, Etude Expérimentale de l'Intelligence. Schleicher Frères & Cie, Paris (1903)
4. Brooks, R.: Intelligence without representation. Artificial Intelligence 47, 139–159 (1991)
5. Charette, R.: WellPoint Hires IBM's "Dr." Watson (April 11, 2013), `http://spectrum.ieee.org/riskfactor/biomedical/diagnostics/wellpoint-hires-ibms-dr-watson`
6. Charmaz, K.: Constructing Grounded Theory. Sage, London (2006)
7. Cohen, P.: If Not Turing's Test, Then What? AI Magazine 26(4), 61–67 (2005)
8. Flanagan, J.: The critical incident technique. Psychological Bulletin 51(2), 327–358 (1954)
9. Fodor, J.: The Language of Thought. Thomas Crowell, London (1975)
10. French, R.M.: Subcognition and the limits of the Turing Test. Mind 99, 53–65 (1990)
11. Hodges, A.: Alan Turing: the Enigma. Burnett, London (1983)
12. Hung, V., Elvir, M., Gonzalez, A.J., DeMara, R.F.: A Method For Evaluating Naturalness in Conversational Dialog Systems. In: Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2647–2652 (2009)
13. ISO/IEC, Ergonomic Requirements for Office Work with Visual Display Terminals (VDT)s - Part 11 Guidance on Usability. ISO-IEC, Geneva (1998)
14. Kirakowski, J., O'Donnell, P., Yiu, A.: Establishing the hallmarks of a convincing chatbot-human dialogue. In: Maurtua, I. (ed.) Human- Computer Interaction, In-Teh., Vukovar (2009)

15. Rzepka, R., Ge, Y., Araki, K.: Naturalness of an utterance based on the automatically retrieved commonsense. In: Proceedings of the Nineteenth IJCAI (2005)
16. Semeraro, G., Andersen, H.H.K., Andersen, V., Lops, P., Abbattista, F.: Evaluation and validation of a conversational agent embodied in a bookstore. In: Carbonell, N., Stephanidis, C. (eds.) UI4ALL 2002. LNCS, vol. 2615, pp. 360–371. Springer, Heidelberg (2003)
17. Shawar, B.A., Atwell, E.: Chatbots: Are they Really Useful? LDV Forum 22 (1), 29–49 (2007)
18. Shieber, S.M.: Does the Turing Test Demonstrate Intelligence or Not? Proceedings of the 21st National Conference on Artificial Intelligence 2, 1539–1542 (2006)
19. Turing, A.: Computing machinery and intelligence. Mind 59, 433–460 (1950)
20. Weizenbaum, J.: ELIZA – A Computer Program for the Study of Natural Language Communication Between Man And Machine. Communications of the ACM 9(1), 36–45 (1966)