The Confederate Effect in Human-Machine Textual Interaction

HUMA SHAH, ODETTE HENRY
School of Computer Science
University of Westminster
Watford Road, Northwick Park, Harrow, HA1 3TP
UNITED KINGDOM

Abstract: - This paper presents a pilot study involving 99 participants analysing conversations between Judges 2, 4 and 7, hidden humans (Confederates), and Jabberwock, bronze prize winner for most human-like machine from Loebner's 2003 Contest, instantiation of Turing's Test for machine intelligence. The transcripts from these conversations were given to children (aged between 8 and 12), and adults (aged between 18-35). The machine was identified in its conversation with Judge 7, but the Confederate Effect featured in the decisions regarding the nature of the Judges and Confederates, who were both sometimes considered machine-like from their textual discourse. Designers of Jabberwock-type programmes may find results presented here useful in improving linguistic productivity of their human-machine textual interaction systems, currently deployed on IKEA's internet site and e-bank Cahoot's web page.

Key-Words: - Confederate Effect, human-machine textual interaction, linguistic productivity, Turing Test

1. Introduction

Loebner Contests [1] provide an annual instantiation of Alan Turing's imitation game [2]. They serve as a platform for Turing's operational test for machine intelligence measured through entries' linguistic productivity (*li.p*). The 2003 Contest exhibited both the Eliza Effect [3]: tendency to accept computer responses as more intelligent than they really are; and the Confederate Effect: where a human's textual discourse is considered machine-like.

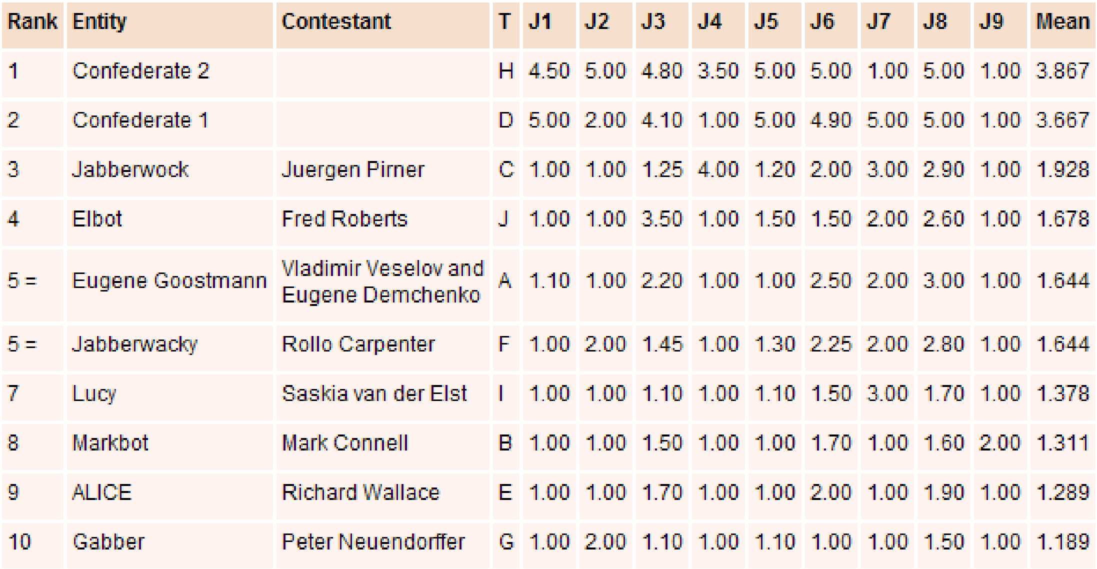
The Confederate Effect became known in 1991 during Loebner's very first realisation of a restricted form of the Turing Test. A hidden human's (Confederate) discourse, limited to five minutes, displayed expertise on the topic of Shakespeare, and was thus considered too knowledgeable to be a human. In a previous study, analysing Loebner machine entrant Mabel [4], the author's discourse was considered machine-like, in comparison with Mabel's [5].

In Loebner's unrestricted 2003 Contest, two hidden humans, the Confederates, and eight machine programmes each chatted unseen to nine Judges for five minutes [6]. The Judges were informed that at least one human, and at least one machine, was present. Table 1 shows the rankings as rated by each of the Judges using the scoring system shown in Table 2. Both Confederates' mean score was less than 4.00 - "probably a human" (see Tables 1 & 2), the score awarded by Judge 4 to Jabberwock (see Table 1), the bronze prize-winner for most human-like machine. The female Confederate 2 topped the rankings with a mean score slightly higher than that of male Confederate 1.

This paper presents a pilot study with 99 participants, 46 children aged 8-12, and 53 adults, aged 18-35. The aim was to find if the participants could discern between human and machine from Loebner 2003 conversations. Section 2 details the methodology adopted, and results. What emerges from the findings is the existence of the Confederate Effect, and, that children as young as eight have knowledge of how human conversation works. Further tests are being evaluated at the time of writing. Nonetheless, designers of Jabberwock-type programmes may want to use these findings to improve *li.p* in their systems.

2. Method

The rationale for choice of Transcripts from Loebner's 2003 ninety conversations was to select those conversations where a Confederate had scored 2.00 or less, rated "probably", or "definitely a machine" from a Judge (see Tables 1 & 2), and where Jabberwock scored 4.00 or higher, rated "probably" or "definitely a human" by a Judge (see Table 2). On that basis four conversations were selected for participants'

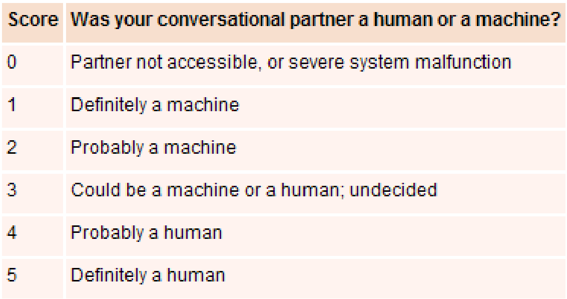


Rank	Entity	Contestant	T	J1	J2	J3	J4	J5	J6	J7	J8	J9	Mean
1	Confederate 2		H	4.50	5.00	4.80	3.50	5.00	5.00	1.00	5.00	1.00	3.867
2	Confederate 1		D	5.00	2.00	4.10	1.00	5.00	4.90	5.00	5.00	1.00	3.667
3	Jabberwock	Juergen Pirner	C	1.00	1.00	1.25	4.00	1.20	2.00	3.00	2.90	1.00	1.928
4	Elbot	Fred Roberts	J	1.00	1.00	3.50	1.00	1.50	1.50	2.00	2.60	1.00	1.678
5 =	Eugene Goostmann	Vladimir Veselov and Eugene Demchenko	A	1.10	1.00	2.20	1.00	1.00	2.50	2.00	3.00	1.00	1.644
5 =	Jabberwacky	Rollo Carpenter	F	1.00	2.00	1.45	1.00	1.30	2.25	2.00	2.80	1.00	1.644
7	Lucy	Saskia van der Elst	I	1.00	1.00	1.10	1.00	1.10	1.50	3.00	1.70	1.00	1.378
8	Markbot	Mark Connell	B	1.00	1.00	1.50	1.00	1.00	1.70	1.00	1.60	2.00	1.311
9	ALICE	Richard Wallace	E	1.00	1.00	1.70	1.00	1.00	2.00	1.00	1.90	1.00	1.289
10	Gabber	Peter Neuendorffer	G	1.00	2.00	1.10	1.00	1.10	1.00	1.00	1.50	1.00	1.189

Table 1: Loebner 2003 Rankings

testing: Judge 2 with Confederate 1 (latter rated probably a machine); Judge 4 with Confederate 1 (latter rated definitely a machine); Judge 4 with Jabberwock (machine rated probably a human), and Judge 7 with Confederate 2 (latter rated definitely a machine), see Table 1.

Although Loebner's 2003 Contest featured male and female judges of varying ages, the organisers did not keep information on which was which. Colby et al's [7] paradigm was adopted in this study, with each participant given just one transcript of a conversation for them to analyse.



Score	Was your conversational partner a human or a machine?
0	Partner not accessible, or severe system malfunction
1	Definitely a machine
2	Probably a machine
3	Could be a machine or a human; undecided
4	Probably a human
5	Definitely a human

Table 2: Loebner 2003 scores

2.1 Study Population

99 participants were recruited including 37 females and 62 males. 46 were children (aged between 8 and 12), of which 22 were female and 24 were male. 53 adults (aged between 18 and 35) included 15 females and 38 males. The children were tested while attending Saturday morning classes at a school in Wembley, UK. The adults were first year undergraduates, participating during Tutorials at the University of Westminster's School of Computer Science, Harrow UK.

2.2 Procedure

Loebner 2003 transcripts were altered so that the names of original conversants would not be revealed. For instance, in the conversation between Confederate 1 (CHH1) and Judge 2, their names were replaced with C1 & C2 respectively. All testing was done in classroom settings. Before transcripts were given to each participant, the researcher engaged them in a discussion about the Hollywood movie "I-Robot" [8].

Participants were asked if they had seen the movie, and if so, had they noticed anything unusual about it. The children's responses included mentioning the high-tech car featured in the movie, and that it was about the future. One female adult mentioned that the robot protagonist expressed emotions during the movie. No child or adult considered it unusual that the robot talked and understood natural language. This suggests that participants may be used to the idea of robots and machines interacting with humans using human languages. Each was given one transcript from the four selected Loebner 2003 conversations.

The Transcripts included written instructions, but these were verbally read out to them at the start. Child participants were asked to detail their age, gender and date/time, and any other language spoken (see Box 1). Adults were asked their age group (18-24; 25-35; 35 & over); the question of other spoken languages was replaced with a question asking if the adults had heard of the Turing Test.

Reference (to be completed by researcher)
Gender:
Age:
Date & time:
Do you speak any language other than English?

Box 1: Child Participants requested details

The participants were given twenty minutes to read their transcript. After reading and analysing the conversations, participants were asked to complete a box regarding what they felt was the nature of the two chatting agents (see sample Box 2).

Agent:	Result: Human	Result: Machine	Gender?	Comments (you may continue at the back of these sheets)
T1				
T2				

Box 2: Participants decision & comments

2.3 Results

The results for each of the four transcripts are detailed in sections 2.3.1 to 2.3.4. Both the Eliza and the Confederate Effect feature in the participants' decisions: Jabberwock, in its conversation with Judge 4, was deemed human by one adult participant. Both Confederates and Judges were considered machine-like by some participants, based on their *li.p* during conversation. In addition, the female Confederate was considered male through her textual discourse by all four children who recognised her as human, and by three adults in Transcript 3: CHH2 conversation with Judge 7. The male Confederate was confused as female by 3 adult participants analysing his conversation with Judge 2 in Transcript 1, and by three children and two adults in Transcript 2 with Judge 4.

2.3.1 Transcript 1: J2 – CHH1

31 subjects read this transcript. Of these, 61% (19) rated CHH1 – male Confederate, a machine. Judge 2 (J2) had given male Confederate (CHH1) a score of 2.00 = "probably a machine". Of the 13 children: 6 girls and 7 boys aged between 8 ½ and 12, 69% (9) considered CHH1 machine, whereas 55.56% (10) adults considered CHH1 machine (see Chart 1). Overall 14 subjects (45%) rated J2's language machine-like, of these four were children and 10 were adults.

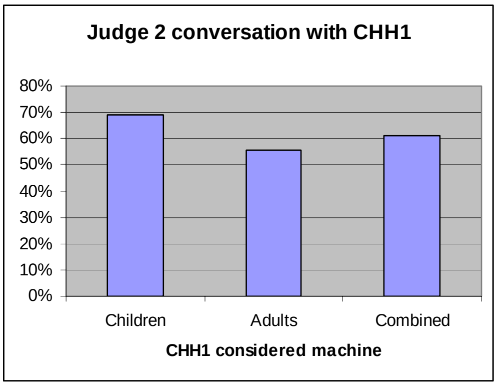


Chart 1: CHH1 = machine

2.3.2 Transcript 2: J4 – CHH1

25 participated (5 females, 20 males) in analysing this conversation. In this transcript CHH1 was named P1, and Judge 4 was named P2. Only 4 subjects (16%) agreed with Judge 4 considering CHH1 a machine, of these 3 were children and 1 an adult. One adult remained undecided on the nature of CHH1.

On closer inspection, all four participants attributing machine-ness to CHH1 were male (three aged 11, and one aged 20). Their comments included: " I think P1 is a machine because he knows a lot" (subject: Loebner child 2003 /14); "P1 robot because it sounds really intelligent" (subject: Loebner child 2003 /21); "sentence 5 – I was asked and you? by P1 means ordered to do this" (subject: Loebner child 2003 /22 – here the subject is referring to transcript sentence number 5, see Table 3); the adult male subject (Loebner Adult 2003/30) gave transcript sentence numbers 7, 13, 15, 18, 19, 21 & 23 from CHH1 as evidence for their machine-ness, (see Table 3).

However, 76% of the participants (19 of 25) analysing this conversation felt that Judge 4 was a machine. Of those, 11 were children (4 females and 7 males) and 8 were adults (1 female and 7 males).

4. P2: Well J how did you get into this?
5. P1: I was asked. And you?
6. P2: I was volunteered.
7. P1: Ah. Ray, may I ask you: are you a computer?
11. P1: This is getting a bit heavy. How about if I ask you how you got here today?
12. P2: By car
13. P1: Idealogically unsound person! What's wrong with the train?
14. P2: It does not run directly.
15. P1: So, what's wrong with getting your bike out?
16. P2: Too far.
17. P1: There are some lovely buses going your way.
18. P2: What do you know about the locality.
19. P1: Quite a bit. I lived here for 17 years.
20. P2: So which bus would you get from heatherside to surrey university
21. P1: Any old bus. Which bus would you get?
22. P2: I would not and anyway certainly not an old one.
23. P1: Goodoh. Let us instead talk about something other than buses. Do you like gardening?

Table 3: segment of CHH1/P1 – J4/P2 conversation

2.3.3 Transcript 3: Judge 7 – CHH2

19 subjects participated reading this transcript. Overall 52.6% (10) agreed with Judge 7 that CHH2 (renamed T1 in the transcript) was a machine. Of these 10, 4 were children and 6 were adults, 9 male and one a female. CHH2, the female Confederate confused the participants with what were considered unusual responses, especially at line 25 (see Table 4) where CHH2 replies “words” to Judge 7 (T2’s) question at line 24 (see Table 4).

Regarding Judge 7, 66.67% (6 out of 9) children participating in this transcript felt this entity was a machine, while 50% of adults did, with 1 of the 10 adults unsure about nature of Judge 7.

2.3.4 Transcript 4: J4 – Jabberwock

24 participants – 12 children and 12 adults read this transcript between Judge 4 and Jabberwock, the bronze prize-winning machine in Loebner’s 2003 Contest, for most human-like machine. Overall 87% (21 participants) recognised Jabberwock as the machine (see Chart 2). Of the 12 children, all (100%) and 9 (75%) of adults identified Jabberwock’s artificial conversation; 2 adults were undecided on Jabberwock’s nature, and one female considered it a human.

Judge 4 had awarded a score of 4.00 – “probably a human” (see Tables 1 & 2) to Jabberwock. Judge 4 was recognised as human by all 12 children (100%), while 7 adults (58.33%) considered Judge 4’s conversation human-like.

16. 2003-10-18-14-19-21 **T1**: Which do you prefer - Brighton or Worthing?
17. 2003-10-18-14-19-42 **T1**: Hello. Who are you?
18. 2003-10-18-14-19-55 **T2**: why should i answer that question?
19. 2003-10-18-14-20-11 **T1**: Hello. Who are you?
20. 2003-10-18-14-20-29 **T2**: you are very pushy!!
21. 2003-10-18-14-20-50 **T1**: Just assertive
22. 2003-10-18-14-21-07 **T2**: that makes sense - why are you assertive
23. 2003-10-18-14-21-24 **T1**: To mask my insecurities
24. 2003-10-18-14-21-43 **T2**: what are your areas of personal confidence?
25. 2003-10-18-14-22-18 **T1**: Words.
26. 2003-10-18-14-22-36 **T2**: how did you acquire confidence in words?
27. 2003-10-18-14-22-53 **T1**: Tell me your name first.
28. 2003-10-18-14-23-01 **T2**: roger
29. 2003-10-18-14-23-22 **T1**: That's nice. Real masculine...
30. 2003-10-18-14-24-07 **T2**: good bye
31. 2003-10-18-14-24-13 **T1**: Bye

Table 4: segment from CHH2/ T1 – J7/T2 conversation

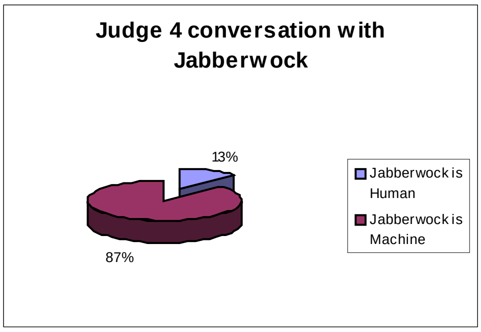


Chart 2: Jabberwock = machine

3. Discussion

The goal of the Loebner Contest is to provide a platform for an operational test for machine intelligence, as devised by Alan Turing. Described in the latter’s seminal 1950 piece, that an unseen machine imitating human textual discourse could be considered intelligent, if it deceived “an average interlocutor after five minutes of questioning” [2] that they were in conversation with another human. Turing did not offer any method for how such a machine could be constructed; this has been left for creators of chatbots – artificial chatting agents such as Jabberwock.

Weizenbaum's Eliza [9] accomplished this feat of deception in 1966, but no one seriously considered it intelligent. As French [10] points out: the very capacity of the Turing Test to probe the deepest, most essential areas of human cognition makes it virtually useless as a real test for intelligence. Since Eliza, chatbots have increased in technical sophistication and capabilities [11] offering textual interaction with humans, but it is how the humans behave conversationally in the Loebner Contests that is far more enlightening.

The rankings in Loebner's 2003 Contest reveal that both Confederates scored higher than the winning machine (female Confederate scored highest at 3.867 compared to the male Confederate mean of 3.667), but both their mean scores were less than 4.00, "probably a human". This may suggest that humans constrain their conversation during the artificial setting of a Loebner Contest, to produce machine-like responses.

This pilot study shows the Eliza effect: attributing intelligence where it does not exist, affected a female adult participant who deemed the bronze-prize wining machine human. Judge 4 had rated Jabberwock as "probably human" (see Tables 1 & 2). The study also reveals the existence of the Confederate Effect: both female and male hidden humans in Loebner 2003 were sometimes considered machine-like from their conversation. Additionally, the three Judges (J2, J4 & J7) in our tests were also considered machine-like by some of the participants.

Results presented in Transcript 1 -section 2.3.1, show that 61% of participants agreed with Judge 2 that Confederate 1 was machine-like, from their textual conversation. In Transcript 3 tests, 52.6% agreed with Judge 7 that Confederate 2 was a machine. The largest variation between our tests and the Loebner Contest results can be seen in Transcript 2, section 2.3.2 and Transcript 4, section 2.3.4. Our results show that Confederate 1's conversation was deemed human by 21 of the 25 participants engaged in analysing the conversation with Judge 4; Judge 4 had awarded Confederate 1 with a score of 1.00 (definitely a machine). In contrast to decision of Judge 4, adult participants in our study, from its conversation with Judge 4, identified Jabberwock as the machine 75 % of the time. Children identified the machine 100% of the time. However, our participants found Judge 4 to be machine-like.

Future work includes evaluating linguistic productivity (*li.p*) of real-world application of Jabberwock systems, for example Kiwilogic's lingubots [12]. Lingubots, for example Anna, is used as a virtual customer service agent by Swedish furniture company IKEA on its web site; another "any questions" query system is used by Cahoot a UK Internet bank.

4. Conclusion

The setting for the Loebner Contests may contribute to boredom, or tiredness in both Confederates involved as hidden humans, and Judges attempting to distinguish between human and machine, from their text-based *li.p*. These factors may result in constrained text-based discourse that appears less human-like causing the Confederate Effect. The findings presented here could prove valuable to designers of human-machine textual interactive systems, especially in single, specialised domains, such as those deployed in e-banking query systems, e.g. Cahoot [13]. The linguistic productivity of these systems could improve by considering the work undertaken in this study.

References:

[1] H. Loebner, Loebner Prize Home Page http://www.loebner.net/Prizef/loebner-prize.html 2003

[2] A.M Turing, Computing Machinery & Intelligence, *Mind*, Vol 59, 1950, pp 433-460

[3] S. Turkle *Life on the screen- Identity in the Age of the Internet*, Phoenix Paperback: London, 1997

[4] D. Hamill, Mabel chatbot at http://www.maybot.com/ 2001

[5] H. Shah Analysing Intelligence-Linguistic Productivity in a Modern Eliza (not published) 2005

[6] University of Surrey's Loebner 2003 Contest: http://www.surrey.ac.uk/dwrc/loebner/results.html

[7] K. M. Colby, S. Weber & F.D. Hilf. Artificial Paranoia, *Artificial Intelligence*, Volume 2, 1971, pp 1-25

[8] I-Robot http://www.irobotmovie.com/ 20th Century Fox 2004

[9] J. Weizenbaum, Eliza - a computer programme for the study of natural language. *Communication of the ACM* Vol. 9 No. 1 1966

[10] R.M. French. Subcognition and the Limits of the Turing Test, *Mind* Vol. 99 No. 393 1990, pp 53.65

[11] K.R Stephens. What has the Loebner Contest told us about Conversant Systems? *Cambridge Center for Behavioral Studies* 2004.

[12] Kiwilogic. Virtual Service Agents at: http://www.kiwilogic.com/?menue_id=10067&submenue_id=10107&id 2005

[13] Cahoot on-line bank “any question” system at: http://www.cahoot.com/ 2005

Tables:

Tables 1 & 2 from University of Surrey’s Loebner 2003 Contest site: http://www.surrey.ac.uk/dwrc/loebner/results.html

Acknowledgements: The authors would like to thank Dr Hugh Loebner for allowing the use of Loebner 2003 transcripts; Dr Evans – Deputy Headmaster, and children of the Copland School in Wembley Middlesex; Undergraduates in Human-Computer Interaction, Introduction to Linguistics for Cognitive Science and AI Fundamentals for participating in this study. Also members of the ROBITRON discussion group on Natural Language Processing, Artificial Intelligence and chatterbots. This discussion group can be found at: http://groups.yahoo.com/group/Robitron/