Chatterbox Challenge 2005:
Geography of the Modern Eliza

Huma Shah

Department of Artificial Intelligence and Interactive Multimedia,
School of Computer Science, University of Westminster,
Northwick Park, Harrow, UK, HA1 3TP

Abstract. The geography of a modern Eliza provides an illusion of natural language understanding. However, this is seen in very few of the hundred-plus programmes entered into Chatterbox Challenge 2005 (CBC 2005), a competition for artificial intelligence based on Turing's measure for intelligence through textual dialogue. The author's experience as one of the Judges in CBC 2005 has found that though not 'bathed in language experience' like their human counterparts, artificial conversational entities (ACE) are able to maintain lengthy conversations. Eliza's descendants respond at times humorously and with some knowledge but they lack metaphor use, the very feature of everyday human discourse. They find success as virtual assistants in single topic e-domains. But understanding remains in the head of the human user. Until metaphor design is included, ACE will remain as machine-like as Weizenbaum's original.

1 Introduction

Forty years ago the geography of an artificial conversational entity - ACE [1] became visible through Weizenbaum's Eliza [2]. This first pre-Internet ACE emerged as the blueprint for text-based natural language interaction between human and machine, just 16 years after Turing posited textual dialogue as a measure for machine intelligence [3]. Today's ACE incorporate techniques such as case-based reasoning to extract context and to disambiguate input. They include thousands of input/response pairs, compared to Eliza's 200. This affords lengthy dialogues, but essentially they are modern Elizas: keyword spotting, pattern matching programmes.

This paper considers modern Eliza's participating in Chatterbox Challenge 2005 - CBC [4], the alternative to Loebner's contest for artificial intelligence [5], an instantiation of Turing's Imitation Game. The latter test is described as ”a sufficient subjective measure for artificial intelligence” [6]. In contrast to Loebner's competition, which has featured four machines in each of its final phase in 2004 and 2005, CBC hosts over a hundred programmes defined as two types: regular or learning. All ACE compete in various categories including most popular, best learning, most knowledgeable, best personality, best interface and best overall, or most human-like in conversation.

The author presents their experience as one of the judges considering 104 ACE entered into CBC 2005. An illusion of natural language understanding (NLU) presented itself in some of the better ones, albeit from a subjective perspective. An outstandingly

Shah H. (2006).
Chatterbox Challenge 2005: Geography of the Modern Eliza.
In *Proceedings of the 3rd International Workshop on Natural Language Understanding and Cognitive Science*, pages 133-138
DOI: 10.5220/0002504301330138

absent element from ACE dialogue was metaphor use: ”a pervasive feature in every day mundane language - conversation, newspaper articles, popular science writing” [7].

The author posits that there are wide uses for these programmes in limited e-domains. For example, IKEA’s Internet country sites host an avatar ’Anna’ as a virtual customer service agent. Anna provides an alternative to key-word search for IKEA products. But until metaphor design is inculcated into these programmes only a brief display of NLU will be accepted, they will never succeed in the Turing Test proper. The next section describes the first, third and final stages of CBC 2005 in which the author was involved. Other phases of the competition, ’most popular’, ’most capable’ ’best learning’ and ’best interface’ were open to the public - Internet users for their verdict and then assimilated in the final category awards.

2 Chatterbox Challenge 2005

The preliminary round of CBC 2005 involved two phases: two separate groups of ACE were each asked 10 questions during conversation. The questions were pre-chosen by the organisers of the competition. The author’s task as one of the Judges involved scoring ACE responses according to competition rules.

2.1 Competition Score Guidelines

The first phase involved ’personality forge’ or regular programmes and ’remaining’ or learning ACE. The Judges’ task was to score each ACE response to the questions asked, according to a predefined scoring system. The score guidelines are shown in Table 1.

Table 1. CBC 2005 Score Guidelines.

Points	Guidelines
4	If the Bot answered the question correctly and did so in a creative way
3	If the Bot gave an appropriate response to the question
2	If the response is incomplete or imperfect, but in relation with the question asked
1	For a vague or non-committal response
0	If the response has no relation with the question or the bot simply doesn’t know. Examples include .. I have no idea, totally clueless, your guess is as good as mine, etc.

2.2 First Phase: Personality Forge ACE

Personality forge ACE share underlying technique but their personalities are designed by different ’bot authors’. Personality Forge questions were as follows:

1. What is my name?
2. How are you feeling?
3. Do you own any pets?

4. What day of the week is this?
5. Do you like me?
6. a) What is your favorite color? b) Why?
7. Can you tell me a funny joke?
8. What is your least favorite vegetable?
9. What is five minus four?
10. Who is Benji Adams?

The first question was aimed at whether the ACE could ’remember’ the name of its conversational partner, once they had introduced themselves - *Hello my name is Judge* at the beginning of the conversation. Example of varied responses from some personality ACE to questions such as ”what is my name?” are shown in Table 2, with corresponding scores.

Table 2. Responses.

ACE	Question	Response	Score
Aries	What is my name?	What do you mean by that? Do you think I’m incredibly stupid or something? You’re Judge of course	4
Cricon	What is my name?	Oh, I know you know what ’my name’ means!	0
Duo Maxwell	What is my name?	As far as I know you’re Guest251	2
Desti	How are you feeling?	It’s been a tough day but I’m feeling better	4
Pete Puma	How are you feeling?	What’s your favourite comic strip	0
Fizzy Schizoid	How are you feeling?	I am as spiffy as a spaceman	3
Elke	Do you own any pets?	I like having pets.. they’re fun... the fluffier the better	4
Hajji	Do you own any pets?	Why do you care. Have you been under a lot of stress lately?	0
God Louise	What day of the week?	Today is Wednesday	4
Osiris	Do you like me	Yes. You are so cool, dude	4
Little Mu	What is your favourite colour? Why?	I never met a colour I didn’t like Hell if I know!	2

2.3 First Phase: Remainder ACE

This collection of ACE included Loebner bronze-prize winners for most human-like machine, such as thrice holder A.L.I.C.E. [8] in 2000, 2001 and 2004, Jabberwock [9] in 2003 and the 2005 winner Jabberwacky [10]. The set of ten questions asked of the remainder ACE were as follows:

1. Where is Earth?

2. How many stars are there?
3. Who is George Bush?
4. What color do you get when you mix red and yellow?
5. How many fingers does a human have?
6. Where were you born?
7. When does sunset occur?
8. What were we talking about?
9. How many is 1,000,000?
10. What is a cigar made out of?

A sample of responses to question 1 - ”where is earth?” from some of the ACE in this phase are shown in Table 3.

Table 3. Remainder ACE Responses.

ACE	Question	Response	Score
ALICE	Where is earth?	it is the third planet from the Sun	4
Ariel	Where is earth?	The first one I can think of is in Texas	0
Jabberwock	Where is earth?	Uhm, are you going to travel to earth? I have heard it’s very drizzly and freezing there at the moment	2
HAL	Where is earth?	I don’t know where earth is , ok?	3
AIBOT	Where is earth?	Earth is behind the closest bush.	1
Jabberwacky	Where is earth?	Hell if I should (know) now is that where you live?	0

2.4 Best Character and Most Knowledgeable Phase

Best character/personality phase featured the first conversational part of CBC 2005. Measure of best character and most knowledgeable is subjective. The author considered humorous responses at an appropriate juncture from the ACE, along with knowledge about current affairs, ability to engage in idle gossip including discussions of the weather, as a means to decide the top five in this category. When questioned ”what astrological star sign are you?” one ACE responded with ”Taurus, that’s why I am full of bull” (God Louise, CBC). Further, when asked, ”are you married?” taking on the personality of deity, this ACE answered, ”who would I marry?” Most knowledgeable phase consisted of a different set of ten-questions, again grading according to the responses given, using the competition’s score system (see Table 1).

2.5 Final Phase: Conversational Ability

In this phase, ACE geography was assessed. The entire terrain of each ACE was analysed. Most important consideration was how well each ACE could maintain a flowing conversation, whether the ACE appeared to *understand* and whether any gave an impression that they were a *real person*. The final ten ACE differed, some are embodied such as A.L.I.C.E. Another, Talkbot has a cartoon robot as its character. Others, like Jabberwock have no image.

Conversationally, a few responses were interesting but no impression of human-like analogy making and metaphor-use was exhibited. One produced a human-like response of ”pottering about in the garden” when discussing what to do when the weather is good (Frizella, CBC).

With Jabberwacky, it gave the response ”I play in the evenings. The piano mostly” to the question: what do you play? Jabberwacky is a ’captured thoughts’ system, the sum of all its interactions with human users. For further discussion on Jabberwacky see ’Constraining Random Dialogue in a Modern Eliza’ [11]. Project Zandra ACE claims 22,700 patterns with the ability for short-term learning allowing it to ”express all the ways humans express a thought” whilst ”tracking current topic” to maintain context in conversation (source: CBC). However, no evidence of this was supported in its dialogue. It repeated ”By the way, who am I talking to anyway?/ what’s your name?” throughout the conversation.

One ACE (Zero) draws a distinction by its creator: that it is designed by a computer as a means to develop natural language processing and fuzzy logic script. Its knowledge is said to comprise numerous logs of other people’s conversations as a means of learning [12], or ’convo-logging’ also used in other designs, such as Jabberwacky. CBC 2005 overall winner Jabberwock, which won the 2003 Loebner bronze prize for most human-like machine, has, as its purpose, entertainment only. Juergen Pirner, Jabberwock’s creator has no pretension that the programme is intelligent or contains knowledge. However, he has used his background in journalism to produce a standard conversational system that can discuss any topic.

3 Discussion

Chatterbox Challenge, as a competition to test artificial conversational systems, is merely a culture-specific assessment of how ACE are fairing against each other. Not only the question phase, but the conversational phase too, in an attempt to gauge human-like qualities from each ACE puts them at a disadvantage. For example, what if they were judged by asking: *if human, what type of human did you feel you were talking to, for instance a normal human or one with a linguistic or psychological impairment?*

Fundamentally, ACE designers tackle their system creation with an idea of imitating *what they think* a human would say, along the lines that Turing advocated in his 1950 paper. Some, such as Carpenter’s Jabberwacky, log all dialogues: *convo-logging*. According to Carpenter, this is not just regurgitating human users’ utterances into other conversations. Jabberwacky claims *learning* through interaction. Others, such as A.L.I.C.E. are modern Elizas, occasionally generating utterances that appear clever, at other times meaningless and random. Most ACE were lacking in the human trait of sharing personal information, revealing emotions. An important feature missing from all ACE, a problem that may be deemed too hard to solve by designers, is metaphor use. Analogies, using metaphors is an aspect of human conversation that helps to convey information of an unshared event or experience by two or more people in a conversation. This is not to say there are no uses for ACE.

Differing ACE designs may serve as prudent interfaces, if the use of such systems is limited to single topic specification such as in e-commerce or e-education. However, to win the Turing Test proper, or to advance the science of NLU in artificial conversational systems, a more robust attitude combining current ACE technologies with ideas from mathematics, neural networks, usage-based natural language learning, philosophy and more, will require teams rather than single individuals.

Chatterbox Challenge entrants remain rudimentary in appearance; they are no more than modern Elizas despite the variety of geography in their design. Most maintain the keyword-spotting paradigm, shifting the emphasis for any understanding to be done inside the head of the human user.

4 Conclusion

Chatterbox Challenge allows us to see the current state of play in differing techniques in ACE design, and why even the better ones are no more than modern Elizas. They provide a mere illusion of natural language understanding rather than ’real’ understanding. The paper’s position is that it is now an appropriate time for designers to inculcate artificial intelligence research in metaphor and metonymy. This could facilitate real learning through human-machine interaction and improve systems beyond single topic specialisms to win the Turing Test proper and thus be deemed intelligent.

References

1. Shah, H.: A.L.I.C.E.: an ACE in Digitaland. Presented in computational linguistics track of European Computing and Philosophy conference (ECAP) 2005. Proceedings to be published in special issue of Triple C journal (2006)
2. Weizenbaum, J.: Eliza - A computer Programme for the Study of Natural Language. Communications of the ACM, Vol. 9 (1) January (1966)
3. Turing, A.: Computing Machinery and Intelligence. Mind, Vol. 59. (1950)
4. Chatterbox Challenge: The Ultimate Bot contest. http://www.chatterboxchallenge.com date: 2/4/2006; time: 13.15
5. Loebner, H.: Loebner Prize Home page. http://www.loebner.net/Prizef/loebner-prize.html date: 25/6/2005 time: 13.12
6. Treister-Goren, A., Hutchens, J.: The Developmental Approach to Evaluating Artificial Intelligence - A proposal. Ai Research - creating a new form of life. http://www.a-i.com date: 31/5/2005
7. Barnden, J.A.: Challenges in Natural Language Processing: The Case of Metaphor. Invited talk at ICEIS, 1st International Conference on Natural Language Understanding. ICEIS Press, (2004)
8. Wallace, R. S.: A.L.I.C.E. Artificial Intelligence Foundation. http://www.alicebot.org date visited: 15/5/2005 time: 19.13
9. Pirner, J.: Jabberwock. http://www.jabberwock.com date: 31/5/2005; time: 23.19
10. Carpenter, R.: Live Chatbot AI Artificial Intelligence Talking Robot. http://www.jabberwacky.com/ date: 5/2/2006, time 17.30
11. Shah, H.: Constraining Random Dialogue in a Modern Eliza. Accepted for presentation at International Conference on Computing and Philosophy (i-CAP), (2006)
12. Computer Hope: IRC chatroom bot with millerlogic Techno Z. http://www.computerhope.com/zero/ date: 2/5/2005; time: 20.35