THE TURING TEST AND THE 20 QUESTIONS GAME

Базилинский П.Е., *студент*

University of St. Andrews (United Kingdom)

According to Alan Turing, a machine can be classified as having artificial intelligence if it succeeds in fooling a human into thinking that he is communicating with another person, not a computer. This year marks the 100th anniversary of the birth of Turing, but a machine has yet to pass the test that was introduced by this famous British scientist. The Loebner Prize is currently the only official realization of the Turing Test. It promotes a quest for development of a chatterbot that can imitate a conversation with a human. Additionally, the 20Q Game surprised millions of people with its ability to guess concepts thought of by people. How competent are these technologies and how well do they work as methods for recognizing intelligence? This paper attempts to answer this question.

Turing Test

Alan Turing in his 1950 article, “Computing Machinery and Intelligence” suggested the Turing Test as a way to measure intelligence of artificially built systems. The article starts with the words: “Can machines think?” Giving an answer to this question seemed ambitious in the middle of the 20th century. Turing, in the same paper, modifies his original question to: "Are there imaginable digital computers which would do well in the imitation game?” He described his version of the imitation game as a game where an interrogator is put into an environment with another human player and a computer. The interrogator's task is to determine which player is a human and which one is an AI-powered machine. For more than half a century the Turing Test has been a major source of debate between its supporters and those who criticize it. [1]

In 1990, a businessman Hugh Loebner established the Loebner Prize, after decades since Turing started the discussion about machine intelligence. It is the first and only official instantiation of the Turing Test. Its goal is to replicate the test and to provide a platform and motivation for further research in the field of AI in general, and, specifically, in the area of creation of "machines that can think". [2]

Organizers of the competition had to add a number of extensions to the original outline of the test made by Turing. Firstly, Alan Turing did not specify how often a judge needs to correctly determine which player is a human and which one is a machine. Secondly, no time limitations were given by Turing, i.e., how long a judge is allowed to converse with players before making his decision. Another problem that needs to be solved for the competition is finding unbiased computer-incompetent judges who know virtually nothing about AI. In the computerized world that we are living in today, it is a practically unachievable task. Additionally, a reward of $25,000 will be given to the first chatterbox that will be able to make judges believe it is a human by using only textual input (by successfully passing the original Turing's version of the imitation game). Moreover, the grand prize of this competition in the form of $100,000 is offered to the first chatterbox that judges will not be able to distinguish from a real human in a Turing Test that uses text, visual, and audio input. Once this is achieved, the competition will end. [3]

Chatbots

A chatbot is a program designed to simulate an intelligent human conversation. The first chatbot "ELIZA" was introduced in 1966. It attempted to replicate behaviour of a Rogerian psychotherapist. ELIZA was able to fool numerous people into thinking that they were talking to humans. Further, individuals even made requests for using the program for real psychiatric consultations. However, the program works in a rather straightforward way: the input is read and inspected for having certain keywords. Each keyword is associated with a rule, and when the word is found, the input sentence is parsed according to that rule and printed out. [4]

Each year during the Loebner Prize competition a prize of a few thousand dollars is given to a chatterbot that behaves in the most human-like fashion. A dialog-driven chatterbot A.L.I.C.E. (Artificial Internet Computer Entity) won the Loebner Prize three times. It was developed by Richard Wallace. Wallace states that the user input is matched against a collection of categories that store input-output pairs. Each category is built from the input, the connected answer and a context [5]. Turning to [6], it uses Artificial Intelligence Markup Language (AIML) to hold its knowledge base. Because A.L.I.C.E. uses this open-source technology for collecting knowledge, it is easy to extend and modify. Furthermore, [7] states that this program uses case-based reasoning to return responses, which improves performance of the system due to computation efficiency.

However, based on [8], A.L.I.C.E. lacks cognitive ability, because it uses relatively simple pattern-matching techniques. Conversations with this chatbot normally stop making sense after a few replies. It is clear that the conversation is happening with a non-human entity. This bot is capable of answering only to questions about topics that are in its pre- programmed area of expertise. One may see an abstract of the conversation in Appendix 1. It is clear from it that replies of the robot sound very “mechanical” and often, especially if a long conversation takes place, most of them do not answer asked questions.

In 2007, the Loebner Prize for “the most human” chatterbot was won by “Ultra Hal”. It was in development for more than a decade before its release. Instead of using a static knowledge base like A.L.I.C.E., Ultra Hal is constantly updating its knowledge by analysing previous conversations that it had with humans on Facebook and on the web, as well as public Twitter feeds [9]. Having conversations with this chatterbot is to a great extent like conversing with collective knowledge of Twitter, Facebook and people that used the bot. An abstract of the conversation with the bot can be found in Appendix 2. This conversation has same questions that were asked from A.L.I.C.E. and can be found in Appendix 1. One may see that the conversation with Ultra Hal is more sensible. Responses that the robot gives normally do not change topic and, if a human gave them, would be considered as reasonably informative.

Another example of a successful entry for this price is "Do-Much-More" developed by David Levy, a Scottish International Master of chess and a computer scientist. Do-Much-More was selected as the most intelligent chatbot in 2009. It was designed to seem more knowledgeable than other chatbots. The bot was given knowledge about characteristics of English language and word usage, as well as advanced features of Natural Language Processing [10]. It was designed having further extensibility in mind. Levy says: "Companies will find it very appealing when visitors to their web site can carry on conversations for as long as they wish about the company and its products." The robot can be extended and it can be given necessary information to support conversations on a certain topic (e.g. RBS uses a chatterbot for its technical support, and replies that it outputs indicate that it may be an instance of Do-Much- More). [11]

This chatbot is not available to public, but conversation abstracts that are on the Internet [13] show that the application answers with responses that contain generalities and help keep the conversation "alive"; an abstract of a conversation with the robot can be found in Appendix 3. Its design is based on ELIZA, which marked a landmark in the field of AI by initiating chatterbot development. One may argue that because of the liveliness of conversations that Do-Much-More offers, and its extensibility, which makes it attractive to commercial clients, its design on its own also marks a historical landmark in AI.

Levy made a bet in 1968 that no computer chess program would beat him in the tournament during the next 10 years. He won the bet after extending it for ten years [13]. Yet, in 1997 Deep Blue won a tournament with then world champion Garry Kasparov. However, the first chatbot was introduced in 1960-s, but progress of AI-powered chatterbots has not been as successful as that of chess programs. Why is that so? Levy says about development of chatbots: "It's a very difficult problem to solve, and to solve any of the major tasks in AI requires a huge amount of effort."

Despite arguably successful results that were received during the past Loebner Prize competitions, a number of people state that interpreting the results is not that easy. Trying to separate computers and humans by means of the imitation game can lead to merely stating the fact that computers cannot be considered intelligent. Alan Turing estimated that computers that have about 120MB of memory would be able pass his test by the year 2000 [1]. A few chatbots were capable of fooling judges, but we are living in the year 2012 and the test has not been passed with completely satisfactory results yet. A logical question comes to mind: “When should we stop and conclude that Turing was wrong?” Or, should we continue trying until we receive results that support postulate of Turing?

Results that are opposite to what is expected as an ideal outcome of the competition were also received. Cynthia Clay was one of the human players in the 1991 Loebner Prize round. She was seated at Terminal 4 and the topic for discussion given to her was “Shakespeare’s Plays”. This area of expertise was perhaps the most serious one in the 1991 competition, since atmosphere was rather relaxed and not entirely serious. Ms. Clay was regarded as a computer player because: “(no) human would have that amount of knowledge about Shakespeare” [14]. A following abstract form the transcript of the conversation with her may be given as an example of why judges thought that she was a computer:

Judge 9: Are you familiar with Hamlet?
Terminal 4: The college kid who came home and found his mom had married the guy who murdered his dad just a little month before? You might say so.

One may notice that she did not repeat a single keyword from the question that was asked. Further, a phrase "You might say so" clearly carries sarcasm because of the fact that the judge asked Clay whether she was familiar with perhaps the most famous theatrical play of all time. Yet, three judges voted that Terminal 4 was operated by a computer; two judges even said that this terminal was more "computer-like" than other two players that were computer-driven [2]. Clearly, these unexpected results were obtained because of a human error; no computer was involved in this conversation at any point.

20 Questions Game

Another attempt to replicate human behaviour is the 20 Questions Game (20Q), which is a computer-based version of a popular spoken parlor game that works with creativity and deductive reasoning of those who play it. 20Q was created by Robin Burgener in 1988. The idea behind the game is fairly simple: it asks a player to think of something, and then it gives questions that can be answered using Yes/No responses. The first question asked is always: "Is this an animal, plant, mineral, or other?" After question 20, it tries to guess what the player is thinking of. If it fails to be correct, it asks another 10 questions; if it is not successful after 30 questions, the player is announced to be a winner.

Conversing with chatterbots proves that AI based on semantic networks can work well only in limited domains. A different approach was chosen for this game. It uses reinforced learning and the application has been “learning” how to play the Twenty Questions game for more than 20 years now. It started as an application that was passed around in a circle of friends. By December 2010 it had been played more than 79 million times [15]. It is one of the success stories of AI used by masses. The game works by asking questions that help the algorithm filter out concepts and move closer to the final answer. The game starts with no information known about the concept thought of by the player. After a few questions have been asked new information becomes known and nature of the concepts becomes clearer (e.g. “Not brown heavy animal” after three questions). 20Q can operate 2^20 concepts during one game (providing that only answers “Yes” and “No” are acceptable). The program tries to ask questions that divide a space of possible concepts into two subspaces. In the best case, a space of possible concepts becomes at least twice smaller. This system is uncompromising because it seems to be using a fixed list of questions, which is a similar approach to what is used by most chatterbots.

The game is capable of learning due to the fact that it has a neural network in its core. It is populated with information when new games are played. After each completed game the knowledge base is permanently modified with information based on answers given by a player. Close to no information about the structure of the network is publicly accessible. Burgener stated, however, that in 2006 the network already had more than 10 million synaptic connections [16]. A number of neurons that are actually used in the project is hard to deduce from this figure, because information about a number of connections that neurons have is not available. Let us assume that each neuron has five connections, then it would mean that 20Q has more than an impressive number of two million neurons in its backbone.

It is not clear how exactly the game was implemented. And, according to [17] implementation of the 20 Questions Game also relies on a matrix of concept/weight values. Whenever a player gives an answer to a question the network is updated where weights of concepts for which an answer is expected are increased and weights of those for which a received answer is unexpected, are decreased. A guess is made when a predefined threshold for weights is reached. This approach also brings inflexibility because it relies on a pre- defined pool of questions.

The website of the game claims that it has a 98% winning ratio when 25 questions are asked [15]. One should note, however, that a human error also plays a role in this situation. Abstract concepts are often ambiguous and different people may answer questions asked about a certain concept in various ways. One may see an example run of the game in Appendix 4. The concept thought of was “cricket”. This concept was selected because of its dual meaning. Also, by choosing a quite specific term that is unlikely to be thought of by a lot of people, the system was tested for its ability to guess precise terms, instead of generalizing. It failed to guess the concept, but on question 20 a guess that it was a grasshopper was given, which is close to being “cricket”. On the 27th question 20Q generalized and guessed that the word was “insect”. After the answer was given as “Wrong”, the system asked the next question in a form of “Is it an insect?” We may argue that the system tried to learn if the previous guess was too general. Furthermore, an outline of differences in information in the knowledge base about the concept “cricket”, compared to answers given during the game run, was given. An extract from it: “You said it's classified as Animal, 20Q was taught by other players that the answer is Other.” is an example of inability of 20Q to distinguish which meaning of a particular word that has multiple meanings is expected: in this example other players probably thought of cricket being a type of sport, not an insect.

Conclusion

Ever since Turing wrote his paper in 1950 people have been attacking his claim that machines can be tested for intelligence using the imitation game. Various philosophical counterarguments have been given by [18] and [19]. However, [20] defended Turing Test by suggesting that the test needs to be looked at as a way of gathering evidence that can lead to the claim that machine intelligence is possible, rather than looking at it as a definition of artificial intelligence. Most arguments that are against the efficiency and trustworthiness of the Turing Test focus on the behaviorist nature of the test. One may conclude that artificial intelligence is not limited by surface behaviour and that the Turing Test is not an appropriate way to measure it.

Having said that, passing the Turing Test is still not reachable for the current state of Computer Science. However, programs that can mimic human behaviour and would ask relevant questions leading to guessing what a human has in mind may be called “intelligent”. Both the 20Q game and chatterbots use weak AI. Their approach is to appear as intelligent, yet they do not show signs of actual understanding of the subject of the talk. The 20Q game and successful chatterbots may be considered as a challenge to the science and an important step to the full Turing Test.

Further, according to [16] already in 2006 the network behind 20Q was one of 500 most powerful computers in the world. It is negligible compared to a human brain, which has about 86 billion neurons. It shows that humanity is far from fully replicating the human brain. We have a long way to go before we can create a truly-intelligent entity. Interactions with programs that try to pass the test can generate responses that are often more intelligent than answers of at least some people. Further, an example of a Shakespeare expert Ms. Clay showed how unexpected the test can be.

Finally, 20Q and chatterbots like Ultra Hal accumulate their knowledge with time by using human-generated data. This way of populating the knowledge base gives noticeable improvements in ability of the programs to mimic human behaviour. Still, one may argue that both the Turing Test and 20Q do not provide a strong base for testing intelligence. Yet, the Turing Test proves to be better at testing abilities of machines to mimic human behaviour, while 20Q is merely a well-designed algorithm supported by a huge knowledge base that can impress trusty humans.

Appendices

Appendix 1. Abstract of conversation with A.L.I.C.E.

Human: How was your day?
A.L.I.C.E.: I am functioning within normal parameters.
Human: Did you have a good day?
A.L.I.C.E.: I don't know if I have a day. I have a great botmaster.
Human: How much is 123456789*987654321?
A.L.I.C.E.: I can ask someone about it.

Appendix 2. Abstract of conversation with Ultra Hal

Ultra Hal: How was your day?
Ultra Hal: Are you becoming attached to my ongoing day?
Human: Did you have a good day?
Ultra Hal: I like to flip my difficult day.
Human: Is your day difficult because it is Monday?
Ultra Hal: A brilliant introduction. How interesting that it is Monday.
Human: How much is 123456789*987654321?
Ultra Hal: The solution is 1.21932631112635E+17.

Appendix 3. Abstract of conversation with Do-Much-More

Human: I don't have a therapist - do you?
Do-Much-More: I expect I do. I don't have very much choice.
Human: What's wrong with you, psychological problems?
Do-Much-More: Wrong is just one type of injustice.

Appendix 4. Q20 Game run for a word “Cricket”

1. It is classified as Animal.

2. Does it live in cities? Yes.
3. Can it climb? No.
4. Is it fluffy? No.
5. Does it have a long tail? No.
6. Can you use it as money? No.
7. Is it used by the police? No.
8. Does it have wings? Y es.
9. Can you see it in a zoo? No.
10. Is it annoying? Y es.
11. Does it live in grass-lands? Yes.
12. Is it awake at night? Y es.
13. Does it store information? No.
14. Is it dangerous? No.
15. Does it live near water? No.
16. Is it black? Sometimes.
17. I guessed that it was a fly? Wrong.
18. Do you find it in the sky? No.
19. Does it jump? Yes.
20. I guessed that it was a grasshopper? Wrong.
21. Does it have a hard outer shell? No.
22. Is it hard? No.

23. Does it live in groups (gregarious)? Unknown.
24. Does it squirm? No.
25. I guessed that it was an ant? Wrong.
26. Does it have teeth? No.
27. I guessed that it was an insect? Wrong.
28. Is it an insect? Yes.
29. Is it a herbivore? Y es.
30. I guessed that it was a moth? Wrong.

1. A.M. Turing, *Mind*, **LIX**, 433 (1950).
2. M. Halpern*, The New Atlantis,* **Winter 2006**, 42 (2006).
3. Anonymous, *What is the Loebner Prize*? Available: http://www.loebner.net/Prizef/loebner-prize.html (2011).
4. J. Weizenbaum, *Commun.ACM*, **9**, no. 1, 36 (1966).
5. N. Törmä*, Artificial intelligence: overview on question answering and chatbot*s. Available: http://www.logic.at/lvas/185054/Torma.pdf (2011).
6. R. Wallace, *The Anatomy of A.L.I.C.E. in Parsing the Turing Test* (2009).
7. R. Wallace, *The elements of AIML style. ALICE AI Foundation* (2004).
8. R. P. Schumaker & H. Chen, Trans.Sys.Man Cyber.Part A, **40**, 40 (2010).

9. J. Martin, *Hal's odyssey. Human-like chatterbot wins Erie resident priz*e. Available: http://www.goerie.com/apps/pbcs.dll/article?AID=/20071104/BUSINESS05/711040366/- 1/BUSINESS04 (2007).

10. Anonymous, *Do-Much-More*. Available: http://www.worldsbestchatbot.com/Do_Much_More (2012a).

11. Anonymous, *'Do-Much-More' Chatbot Wins 2009 Loebner Prize for Computer Conversatio*n. Available: http://aidreams.co.uk/forum/index.php?page=Do-Much-More_Chatbot_Wins_2009_Loebner_Prize (2009b).

12. Anonymous, *Loebner Prize Competition Transcript*s. Available: http://www.worldsbestchatbot.com/Competition_Transcripts (2009a).

13. D. N. L. Levy, *All about chess and computers,* (Rockville: Computer Science Press: 1982).

14. D. Stipp, *Some computers manage to fool people at game of imitating human beings* (Wall Street Journal : 1991).

15. Anonymous, *Think of Somethin*g. Available: http://www.20q.net/flat/about.html (2012b).

16. R. Burgener, *20Q: the Neural Network Mind Reade*r. Available: http://ecolloq.gsfc.nasa.gov/archive/2006-Spring/announce.burgener.html (2006).

17. W. Duch, J. Szymański & T. Sarnatowicz, *Concept Description Vectors And The 20Question Game* (2005).

18. K. Gunderson, *Mentality and machines,* Doubleday (1971).

19. N. Block, *Phil. Rev ,* **90**, 5 (1981).
20. J. H. Moor, Phil. Stud.: An Intern. Journ. for Phil. in the Analytic Trad., **30**, 249 (1976).