WINOGRAD SCHEMA CHALLENGE: A STEP BEYOND THE TURING TEST

Domingo Senise de Gracia
MSc in Artificial Intelligence
Universidad Politécnica de Madrid
Boadilla del Monte, 28660 Madrid (Spain)
domingo.senise@haitta.com

ABSTRACT
In this paper I explain the Winograd Schema (WS) Challenge: an alternative to the Turing test -the Imitation Game.

The Winograd Schema Challenge is a test of machine intelligence proposed by Hector Levesque, a computer scientist at the University of Toronto, in 2011. It is a multiple-choice test that involves responding to typed English sentences of a very specific structure: they are instances of what are called Winograd Schemas -named after Terry Winograd, a professor of computer science at Stanford University.

Unlike the Turing test, the subject is not required to engage in a conversation and fool an interrogator into believing s/he is dealing with a person. In fact the nature of the Turing test has come under scrutiny, especially since an AI chatbot named Eugene was claimed to pass it in 2014. The chatbot was not intelligent at all—it's just really good at making you overlook the times when it was stupid, while emphasizing the periodic interactions when its algorithm knew how to answer the questions that you asked it. The WS Challenge was proposed in part to ameliorate these problems and to avoid this kind of situation.

Author Keywords
Turing Test, Winograd Schema Challenge, Terry Winograd, Hector Levesque, ELIZA, Loebner competition, the Imitation Game, Recognizing Textual Entailment (RTE) challenge, Eugene, Natural Language Processing, NLP, Common Sense Reasoning

THE STEP BEYOND

The well-known Turing test was first proposed by Alan Turing (1950) as a practical way to defuse what seemed to him to be a pointless argument about whether or not machines could think. He proposes that, instead of asking such a vague question and then getting caught up in a debate about what it means to really be thinking, we should ask whether a machine would be capable of producing behavior that we would say required thought in people. The sort of behavior he had in mind was participating in a natural conversation in English over a teletype in what he called **the Imitation Game**. The idea, roughly, is that if an interrogator were unable to tell after a long, free flowing and unrestricted conversation with a machine whether s/he was dealing with a person or a machine, then we should be prepared to say that the machine was thinking.

The Turing test does have some troubling aspects, however. Firstly, **the central role of deception**. Consider the case of an intelligent machine trying to pass the test. It must converse with an interrogator and not just show its stuff, but fool her/him into thinking s/he is dealing with a person. To imitate a person well without being evasive, the machine will need to assume a false identity (to answer "How tall are you?" or "Tell me about your parents."). We should much prefer a machine should be able to show us that it is thinking, without having to pretend to be somebody or to have some property (like being tall) that it does not have.

Secondly, we might also question **whether a conversation in English is the right sort of test**. Conversations are so adaptable and can be so wide-ranging, they facilitate deception and trickery. Consider, for example, ELIZA (Weizenbaum 1966) a program using very simple means, was able to fool some people into believing they were conversing with a psychiatrist. The deception works at least in part because we are extremely forgiving in terms of what we will accept as legitimate conversation, and regarding this view think about also the Loebner competition (Shieber 1994), a restricted version of the Turing Test. What is striking about transcripts of conversations during this competition is the fluidity of the responses from the subjects: elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts, points of order. Everything, except clear and direct answers to questions. We agree with Turing that the question is whether or not a certain intelligent behavior can be achieved by a computer program. Nonetheless a free-form conversation as advocated by Turing may not be the best vehicle for a formal test, as it allows a cagey subject to hide behind a smokescreen of playfulness, verbal tricks, and canned

responses. In order to counteract these deception and trickery, one promising proposal was the Recognizing Textual Entailment (RTE) challenge. In this case, a subject is presented with a series of yes-no questions concerning whether one English sentence (A) entails another (B). Two example pairs:

A: Time Warner is the world's largest media and internet company.

B: Time Warner is the world's largest company.

A: Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday.

B: Edvard Munch painted "The Scream."

Getting the correct answers (no and yes above, respectively), clearly requires some thought.

A problem with this challenge, however, is that it rests on the notion of entailment. What if the second (B) above was this:

B: The recovered painting was worth more than $1,000.

Technically, this is not an entailment of (A), although it would certainly be judged true.

As an improvement, the Winograd Schema (or WS) challenge, a variant of the RTE, was developed. It requires subjects to answer binary questions, but without depending on an explicit notion of entailment. The WS Challenge is a small reading comprehension test involving single binary questions. Two examples will illustrate[1]:

The trophy would not fit in the brown suitcase because it was too big.

What was too big?

Answer 0: the trophy

Answer 1: the suitcase

Joan made sure to thank Susan for all the help she had given.

Who had given the help?

Answer 0: Joan

Answer 1: Susan

In each of the questions we have the following four features[1]:

1. Two parties are mentioned in a sentence by noun phrases. They can be two males, two females, two inanimate objects or two groups of people or objects.
2. A pronoun or possessive adjective is used in the sentence in reference to one of the parties, but is also of the right sort for the second party. In the case of males, it is "he/him/his"; for females, it is "she/her/her" for inanimate object it is "it/it/its," and for groups it is "they/them/their."
3. The question involves determining the referent of the pronoun or possessive adjective. Answer 0 is always the first party mentioned in the sentence, and Answer 1 is the second party.
4. There is a word (called the special word) that appears in the sentence and possibly the question. When it is replaced by another word (called the alternate word), everything still makes perfect sense, but the answer changes.

There are no limitations on what the sentences can be about, or what additional noun phrases or pronouns they can include. Ideally, the vocabulary would be restricted enough that even a child would be able to answer the question.

Regarding how it works the fourth feature, consider the first example, the special word is "big" and its alternate is "small;" and in the second example, the special word is "given" and its alternate is "received." These alternate words only show up in alternate versions of the two questions[1]:

The trophy would not fit in the brown suitcase because it was too small.

What was too small?

Answer 0: the trophy

Answer 1: the suitcase

Joan made sure to thank Susan for all the help she had received. Who had received the help?

Answer 0: Joan

Answer 1: Susan

With this fourth feature, we can see that clever tricks involving word order or other features of words or groups of words will not work. The claim is that doing better than guessing requires subjects to figure out what is going on.

The need for thinking is perhaps even more evident in a much more difficult example, a variant of which was first presented by Terry Winograd (1972)[1]:

The town councillors refused to give the angry demonstrators a permit because they feared violence.

Who feared violence?

Answer 0: the town councillors

Answer 1: the angry demonstrators

Here the special word is “feared” and its alternate is “advocated” as in the following:

The town councillors refused to give the angry demonstrators a permit because they advocated violence.

Who advocated violence?

Answer 0: the town councillors

Answer 1: the angry demonstrators

You need to have background knowledge that is not expressed in the words of the sentence to be able to sort out what is going on and decide that it is one group that might be fearful and the other group that might be violent. And it is precisely bringing this background knowledge to bear that we informally call **thinking**.

It is not necessary that the special word and its alternate be opposites (like “big” and “small”). Here are two examples where this is not the case[1]:

Paul tried to call George on the phone, but he was not (). Who was not ()?

Answer 0: Paul

Answer 1: George

Special: successful

Alternate: available

The lawyer asked the witness a question, but he was reluctant to () it.

Who was reluctant?

Answer 0: the lawyer

Answer 1: the witness

Special: repeat

Alternate: answer

Two specific mistakes must be avoided in the WS Challenge[1]:

First mistake: questions whose answers are in a certain sense too obvious.

Consider the following WS:

- The women stopped taking the pills because they were (). Which individuals were ()?

  Answer 0: the women

  Answer 1: the pills

  special: pregnant

  alternate: carcinogenic

In this case, because only the women can be pregnant and only the pills can be carcinogenic, the questions can be answered by ignoring the sentence completely and focussed just on certain keywords.

Second mistake: answers that are not obvious enough.

- Frank was jealous when Bill said that he was the winner of the competition. Who was the winner?

  Answer 0: Frank

  Answer 1: Bill

An alternate word that points to Frank as the obvious winner. Consider this:

- Frank was pleased when Bill said that he was the winner of the competition.

The trouble here is that it is not unreasonable to imagine Frank being pleased because Bill won. The sentence is too ambiguous.

In the end, what a subject will consider to be obvious will depend to a very large extent on what s/he knows. The “easier” questions are not easier because they can be answered in a more superficial way (using, for example, only statistical properties of the individual words). Rather, they differ on the subject's background knowledge assumed. Consider, for example, this intermediate case[1]:

- The large ball crashed right through the table because it was made of (). What was made of ()?

  Answer 0: the ball

  Answer 1: the table

  special: steel

  alternate: styrofoam

For adults who know what styrofoam is, this WS is obvious. But for individuals who may have only heard the word a few times, there could be a problem.

WS challenge does not allow a subject to hide behind a smokescreen of verbal tricks, playfulness, or canned responses, as it might happen with the Turing Test; with a very high probability, anything that answers correctly a series of these questions is thinking in the full-bodied sense we usually reserve for people.

DISCUSSION

AI is more than just technology. **AI is the study of intelligent behavior in computational terms**. The science of AI studies intelligent behavior, not who or what is producing the behavior. It may study natural language understanding, for instance, but not natural language understanders. This is what makes AI quite different from the study of people (in neuroscience, psychology, cognitive science, and so on).

There is a tendency in AI to focus on behavior in a purely statistical sense: can we engineer a system to produce a

desired behavior with no more errors than people would produce?

Unfortunately, this can lead us to systems with very impressive performance that are nonetheless *idiot-savants* [2]. **How can this be intelligence if it is just inserting probabilistic statistical power in the problem and waiting to see what happens, without any underlying knowledge?**

Following the extraordinary idea first proposed by John McCarthy in 1959 [4], we should put aside any idea of tricks and shortcuts, and focus instead on what needs to be *known*, how to represent it symbolically, and how to use the representations to produce an intelligent behavior.

Knowledge is not power if it cannot be acquired in a suitable symbolic form, or if it cannot be applied in a tractable way. These point to two significant hurdles faced by the McCarthy approach:

1. Much of what we come to know about world and the people around us is not from personal experience, but **is due to our use of language.** And yet, it appears that we need to use extensive knowledge to make good sense of all this language.
2. Even the most basic child-level knowledge seems to call upon a wide range of **logical constructs**. And yet, symbolic reasoning over these constructs seems to be much too demanding computationally.

After more than 60 years and several failed trials, we might well wonder if an AI researcher will ever be able to overcome them.

Pressure to obtain results being so hard, it is not surprising that many AI researchers have returned to less radical methods (more mathematics-statistics aligned) to focus on behaviors that are seemingly less knowledge-intensive -*e.g.*, recognizing hand-written digits, following faces in a crowd, and so on. Indeed the results have been amazing.

But these results should not deceive us: **our best behavior does include knowledge-intensive activities such as participating in natural conversations**.

We should carefully study how simple knowledge bases might be used to make sense of the simple language needed, to build slightly more complex knowledge bases, and so on.

We should avoid being deeply convinced by what appears to be the most promising approach of the day. We see this in the fashions of AI research over the years[2]: first, automated theorem proving is going to solve it all; then, the methods appear too weak, and we favor expert systems; then the programs are not situated enough, and we move to behavior-based robotics; then we come to believe that learning from big data is the answer; and on it goes.

It would be much better to admit that, for instance regarding natural language, other AI approaches, less in vogue and implying more hard working, will be needed for dealing with it. This will help AI progress in a steadier and more solid fashion.

Will a computer ever hold a free natural conversation with a human being without cheap tricks?

As always in life, it will depend only and just only on us: on how much perseverance, inventiveness, and wishes of hard working we will bring to the task. By the end of the day, language mastering is not an easy question: we, human beings, have been coping with it for around the last 50,000 years.

A long and exciting challenge lies ahead of us.

CONCLUSION

In this paper I have written about the Winograd Schema Challenge as an improvement of the Turing test, regarding the accuracy of a proof to verify intelligent behavior in an AI machine.

I have explained the structure of the schemas -of which the Challenge consisted, in order to avoid deceitful behavior by non-intelligent bots, and eventually I have pondered about the approach we should follow if eventually we want to include true human-like knowledge into machines in order to solve such a difficult subject as it is language.

APPENDIX

A brief collection of Winograd Schemas[5]:

1. Although they ran at about the same speed, Sue beat Sally because she had such a [good/bad] start. Who had a [good/bad] start?

Answers: Sue/Sally.

2. The sculpture rolled off the shelf because it wasn't [anchored/level]. What wasn't [anchored/level]?

Answers: The sculpture/the shelf.

3. Sam's drawing was hung just above Tina's and it did look much better with another one [below/above] it. Which looked better?

Answers: Sam's drawing/Tina's drawing.

4. Anna did a lot [better/worse] than her good friend Lucy on the test because she had studied so hard. Who studied hard?

Answers: Anna/Lucy

5. The firemen arrived [after/before] the police because they were coming from so far away. Who came from far away?

Answers: The firemen/the police.

6. Frank was upset with Tom because the toaster he had [bought from/sold to] him didn't work. Who had [bought/sold] the toaster?

Answers: Frank/Tom.

7. Jim [yelled at/comforted] Kevin because he was so upset. Who was upset?

Answers: Jim/Kevin.

8. The sack of potatoes had been placed [above/below] the bag of flour, so it had to be moved first. What had to be moved first?

Answers: The sack of potatoes/the bag of flour.

9. Pete envies Martin [because/although] he is very successful. Who is very successful?

Answers: Martin/Pete.

10. I was trying to balance the bottle upside down on the table, but I couldn't do it because it was so [top-heavy/uneven]. What was [top-heavy/uneven]?

Answers: the bottle/the table.

11. I spread the cloth on the table in order to [protect/display] it. To [protect/display] what?

Answers: the table/the cloth.

12. The older students were bullying the younger ones, so we [rescued/punished] them. Whom did we [rescue/punish]?

Answers: The younger students/the older students.

13. I poured water from the bottle into the cup until it was [full/empty]. What was [full/empty]?

Answers: The cup/the bottle.

14. Susan knows all about Ann's personal problems because she is [nosy/indiscreet]. Who is [nosy/indiscreet]?

Answers: Susan/Anne.

REFERENCES

1.Hector Levesque, The Winograd Schema Challenge,. Commonsense, 2011.

2.Hector Levesque, On Our Best Behaviour, IJCAI Research Excellence Award Presentation, 2013.

3.Evan Ackerman, Can Winograd Schemas Replace Turing Test for Defining Human-Level AI? IEEE Spectrum, July 29, 2014.

4.John McCarthy, Programs with Common Sense, 1959

5.A Collection of Winograd Schemas.

6.What is the Winograd Schema Challenge? commonsensereasoning.org