The Winograd Schema Challenge

Hector J. Levesque

Department of Computer Science, University of Toronto

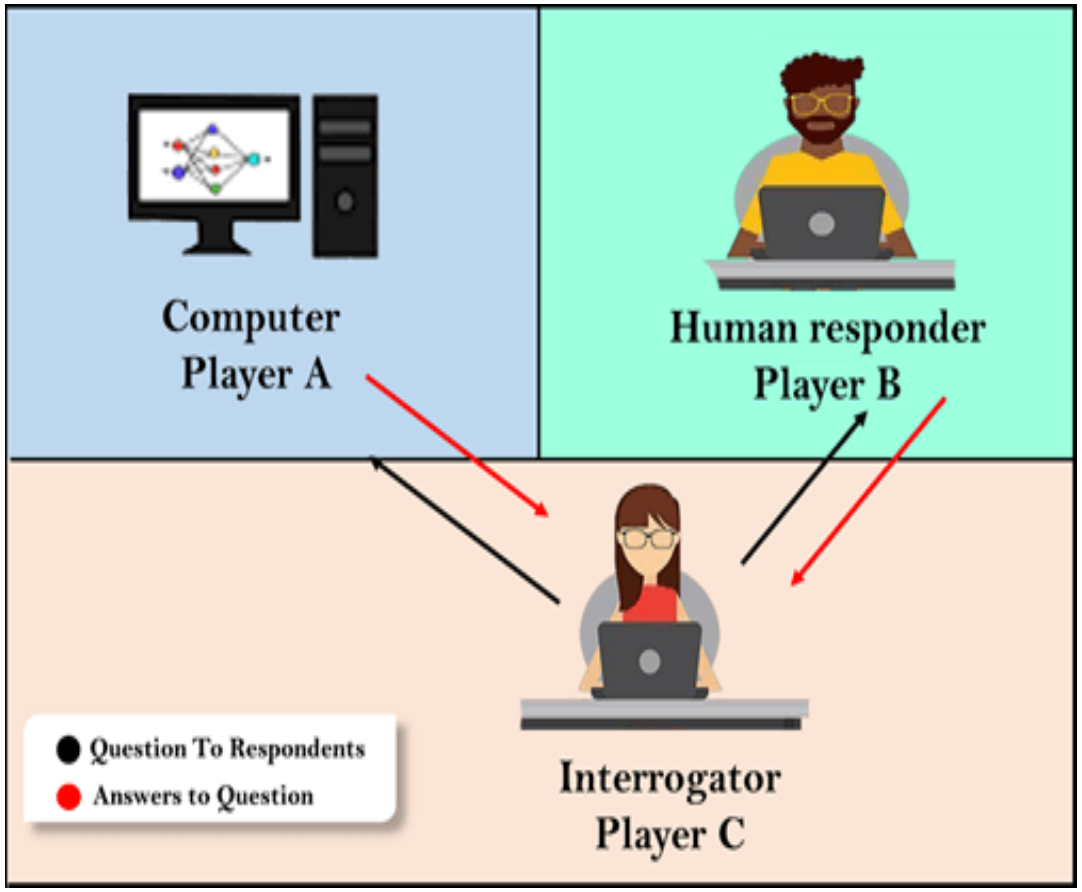
AAAI '11 Spring Symposium (SS-11-06)

Presented by Koo hyeongseok
gudtjrdltka@korea.ac.kr
2021.01.21

How can prove that machines think?

: An alternative to the Turing Test

Turing Test:



- First proposed by Alan Turing in 1950
- To show whether machines could think

Method

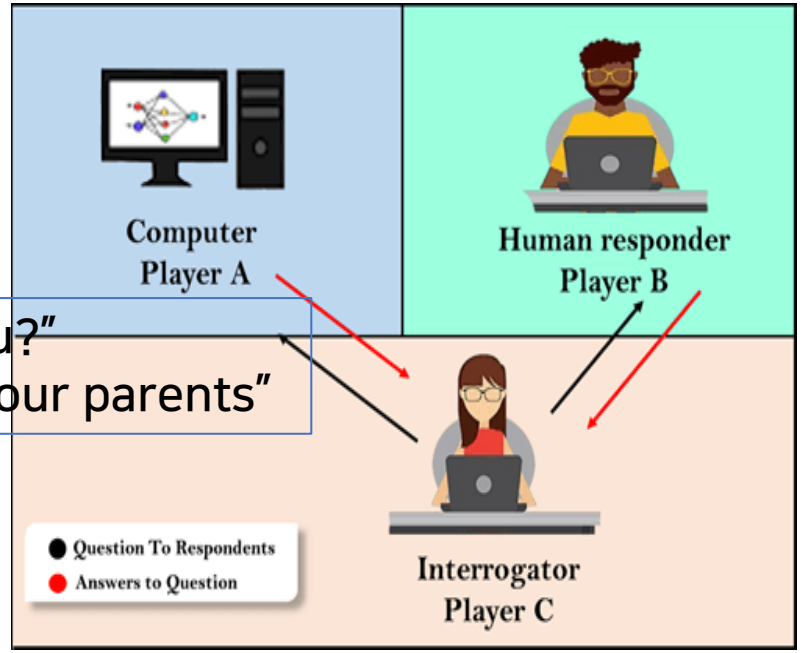
1) A long, free conversation with a machine
2) Then, an interrogator determines whether he/she was dealing with a person or a machine
3) If he/she was unable to tell, we should say that the machine was thinking

Related work

Troubling aspects of Turing Test:

To answer these questions,
the machine needs to assume a false identity

Q: “How tall are you?”
Q: “Tell me about your parents”



It is better if
the machine can show that it is thinking
without pretending to have some property that it does not have

Troubling aspects of Turing Test:

Turing Test: Based on Free form conversations
⇒ Conversations are so adaptable, and facilitate deception and trickery

Ex)

ELIZA (Weizenbaum 1960)
: With very simple means, fool some people into believing they were conversing with a psychiatrist.
: Too restricted case

Loebner Competition (Shieber 1994)
: The end of conversations is always decided

Related work

Desirable features of a new type of Turing Test:

- It involves the subject responding to a broad range of English sentences
- Native English-speaking adults can pass it easily
- It can be administered and graded without expert judges
- When people pass the test, we would say they were thinking

Recognizing Textual Entailment (RTE) challenge

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

=> Does (A) entail (B)? Yes or No

Related work

A problem of RTE challenge:

- It rests on the notion of entailment

A: Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday.
B: The recovered painting was worth more than $1000

Technically, B is not an entailment of A.
However, It would certainly be judged true.

⇒ Propose Winograd Schema Challenge

- A Variant of the RTE challenge
- Not depending on an explicit notion of entailment
- Small reading comprehension test & Answering binary questions

Winograd Schema Challenge

Winograd Schema Challenge

Four features of Winograd Schema Challenge:

1. Two parties are mentioned in a sentence by noun phrases

3. The question involves determining the referent of the pronoun

- 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

2. A pronoun is one of the parties, but is also of the right sort for the second party

Answer 0 is always the first party mentioned in the sentence, and Answer 1 is the second party.

4. There is a special word. When it is replaced by the alternate word, the answer changes

The fourth feature of Winograd Schema Challenge:

How do we know that thinking is required to get a correct answer with high probability?

How do we know that there is not some trick?

⇒ This is where the fourth requirement comes in

- 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

4. There is a special word.
When it is replaced by the alternate word, the answer changes

Winograd Schema Challenge

The fourth feature of Winograd Schema Challenge:

- 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

- Special word => big => Answer 0
- Alternate word => small => Answer 1

Contexts where "big" can appear are statistically quite similar to those where "small" can appear, and yet the answer must change

⇒ Having access to a large corpus of English text would likely not help much

⇒ To solve this question, thinking is needed
⇒ Having and using background knowledge that is not expressed in the words of the sentence

Winograd Schema Challenge

Pitfall 1:

- The women stopped taking the pills because they were < >. Which individuals were < >?

- Answer 0: the women
- Answer 1: the pills

- Special: pregnant
- Alternate: carcinogenic

⇒ Only the women can be pregnant
⇒ Only the pills can be carcinogenic

⇒ The questions can be answered by merely finding the permissible links (learned by sampling a large corpus)

Winograd Schema Challenge

Pitfall 2:

Original version

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

- Answer 0: Frank
- Answer 1: Bill

⇒ Frank being pleased because Bill won / Frank won
⇒ Both sentences are reasonable

⇒ The sentence is too ambiguous

Better version

- Frank felt < > when his longtime rival Bill revealed that he was the winner of the competition. Who was the winner?

Winograd Schema Challenge

Dataset:

150 examples
=> https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

Wsc273 (in tensorflow)
⇒ https://www.tensorflow.org/datasets/catalog/wsc273
⇒ Source code: tfds.text.wsc273.Wsc273

	idx	label	option1	option1_normalized	option2	option2_normalized	pronoun_end	pronoun_start	pronoun_text	text
0	163	1	Fred	Fred	George	George	72	70	he	Fred watched TV while George went out to buy groceries. After an hour he got back.

Discussion and Conclusion

1) Proposing WS challenge as an alternative to the Turing Test

2) It involves responding to typed English sentences, instead of conversation and doesn't need an interrogator.

3) Anything that answers correctly WS questions is thinking
(Whether or not a subject that passes the test is really thinking is the philosophical question that Turing sidesteps)

GLUE SuperGLUE Leaderboard

Rank	Name	Model	URL	Score	WNLI
1	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	94.5
2	HFL iFLYTEK	MacALBERT + DKM		90.7	94.5
3	Alibaba DAMO NLP	StructBERT + TAPT		90.6	94.5
4	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	94.5
5	ERNIE Team - Baidu	ERNIE		90.4	94.5
6	T5 Team - Google	T5		90.3	94.5
7	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	94.5
8	Huawei Noah's Ark Lab	NEZHA-Large		89.8	94.5
9	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	94.5

- The size of the dataset seems too small
- It seems very hard to make new questions
- Only guessing pronouns => Prove the ability of using background knowledge?

감사합니다