The Winograd Schema Challenge

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

Leora Morgenstern
S.A.I.C
Arlington, VA 22203
leora.morgenstern@saic.com

Abstract

In this paper, we present an alternative to the Turing Test that has some conceptual and practical advantages. A Winograd schema is a pair of sentences that differ only in one or two words and that contain a referential ambiguity that is resolved in opposite directions in the two sentences. We have compiled a collection of Winograd schemas, designed so that the correct answer is obvious to the human reader, but cannot easily be found using selectional restrictions or statistical techniques over text corpora. A contestant in the Winograd Schema Challenge is presented with a collection of one sentence from each pair, and required to achieve human-level accuracy in choosing the correct disambiguation.

1 Introduction

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. In a nutshell, 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 focus on *observable behaviour* and ask whether a machine would be capable of producing behaviour that we would say required thought in people. The sort of behaviour he had in mind was participating in a natural conversation in English over a teletype in what he calls 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 she was dealing with a person or a machine, then we should be prepared to say that the machine was thinking. Requiring more of the machine, such that as that it look a certain way, or be biological, or have a certain causal history, is just arbitrary chauvinism.

It is not our intent to defend Turing's argument here (but see the Discussion section below). For our purposes, we simply accept the argument and the emphasis Turing places on intelligent behaviour, counter to critics such as Searle (2008). We also accept that typed English text is a sufficient medium for displaying intelligent behaviour, counter to critics such as Harnad (1989). That is, assuming that *any* sort of behaviour is going to be judged sufficient for showing the presence of thinking (or understanding, or intelligence, or whatever appropriate mental attribute), we assume that typed English text, despite its limitations, will be a rich enough medium.

2 The trouble with Turing

The Turing Test does have some troubling aspects, however. First, note the central role of *deception*. Consider the case of a future intelligent machine trying to pass the test. It must converse with an interrogator and not just show its stuff, but fool her into thinking she is dealing with a *person*. This is just a game, of course, so it's not really lying. But 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."). All other things being equal, we should much prefer a test that did not depend on chicanery of this sort. Or to put it differently, 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.

We might also question whether a *conversation* in English is the right sort of test. Free-form conversations are no doubt the best way to get to know someone, to find out what they think about something, and therefore *that* they are thinking about something. But conversations are so adaptable and can be so wide-ranging that they facilitate deception and trickery.

Consider, for example, ELIZA (Weizenbaum 1966), where a program (usually included as part of the normal Emacs distribution), 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. A Rogerian psychiatrist may say very little except to encourage a patient to keep on talking, but it may be enough, at least for a while.

Consider also the Loebner competition (Shieber 1994), a restricted version of the Turing Test that has attracted considerable publicity. In this case, we have a more balanced conversation taking place than with ELIZA. What is striking about transcripts of these conversations is the fluidity of the responses from the subjects: elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts, points of order. Everything, it would seem, except clear and direct

answers to questions. And how is an interrogator supposed to deal with this evasiveness and determine whether or not there is any real comprehension behind the verbal acrobatics? More conversation. "I'd like to get back to what you said earlier." Short conversations are usually inconclusive; unsurprisingly, the Loebner competition gives judges only 5 minutes to determine whether or not they are conversing with a person or a computer (Christian 2011), not nearly enough to get through the dust cloud of (largely canned) small talk and jokes that the winning programs usually have. Even with long conversations, two interrogators looking at the same transcript may disagree on the final verdict. Grading the test, in other words, is problematic.

How can we steer research in a more constructive direction, away from deception and trickery? One possibility is something like the *captcha* (von Ahn et al. 2003). The idea here is that a distorted image of a multidigit number is presented to a subject who is then required to identify the number. People in general can easily pass the test in seconds, but current computer programs have quite a hard a time of it (cheating aside).[1]

So this test does, at least for now, distinguish people from machines very well. The question is whether this test could play the role of the Turing Test. Passing the test clearly involves some form of cognitive activity in people, but it is doubtful whether it is thinking in the full-bodied sense that Turing had in mind, the touchstone of human-level intelligence. We can imagine a sophisticated automated digit classifier, perhaps one that has learned from an enormous database of distorted digits, doing as well as people on the test. The behaviour of the program may be ideal; but the scope of what we are asking it to do may be too limited to draw a general conclusion.

3 Recognizing Textual Entailment

In general, what we are after is a new type of Turing Test that has these desirable features:

- 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;
- no less than with the original Turing Test, when people pass the test, we would say they were thinking.

One promising proposal is the *recognizing textual entailment* (RTE) challenge (Dagan, Glicksman, and Magnini 2006; Bobrow et al. 2007; Rus et al. 2007). In this case, a subject is presented with a series of yes-no questions concerning whether one English sentence (A), called the *text* (T), *entails* another (B), called the *hypothesis* (H). Two example pairs adapted from (Dagan, Glicksman, and Magnini 2006) illustrate the form:

- 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."

This is on the right track, in our opinion. Getting the correct answers (no and yes above, respectively), clearly requires some thought. Moreover, like the captcha, but unlike the Turing Test, an evasive subject cannot hide behind verbal maneuvers. Also, in terms of a research challenge, incremental progress on the RTE is possible: we can begin with simple lexical analyses of the words in the sentences, and then progress all the way to applying arbitrary amounts of world knowledge to the task.

There are two interrelated problems with this challenge.

First it rests on a somewhat unclear notion of entailment. Of course a precise definition of this concept exists (assuming a precise semantics, like in logic), but subjects could not be expected to know or even understand it. The researchers instead explain to subjects that "T entails H if, typically, a human reading T would infer that H is most likely true" (Dagan, Glicksman, and Magnini 2006). The fact that we need to predict what humans would do, and indeed how they would reason about what is "likely" to be true, which forces the RTE challenge to rest on a nonmonotonic notion of entailment, is troubling. We know, in fact, that what is likely to follow from a set of premises can vary widely, given the semantics of a particular nonmonotonic logic and on the formalization of a background theory (Hanks and McDermott 1987).

Moreover, entailment may not always coincide with human judgement of what is most likely true under certain circumstances. What if the second (B) above was this:

- B: The recovered painting was worth more than $1000.

Technically, this is not an entailment of (A), although it is very likely to be judged true! Of course, subjects can be trained in advance to help sort out issues like this, but it would still be preferable for a practical test not to depend on such a delicate logical concern.

The second problem is perhaps related to the difficulty of getting a firm handle on the more problematic aspects of the RTE notion of entailment: In practice (see (Roemmele, Bejan, and Gordon 2011)), the RTE challenge has focused more on inferences that are necessarily true due to the meaning of the text fragment than on default inferences. This results in a challenge that is easier than might be imagined from the original description of the RTE challenge. Examples of text and hypothesis pairs used in the development set for the 2010 RTE challenge include the following, cited in (Majumdar and Bhattacharyya 2010):

- A: Arabic television Al-Jazeera said Tuesday the kidnappers of a U.S. woman journalist abducted in Baghdad had threatened to kill her if female prisoners in Iraq were not freed within 72 hours .... Al-Jazeera reiterates its rejection and condemnation of all forms of violence targeting journalists and demands the release of the US journalist Jill Carroll, the station said.

[1] Cheating will always be a problem. The story with captchas is that one program was able to decode them by presenting them on a web page as a puzzle to be solved by unwitting third parties before they could gain access to a free porn site! Any test, including anything we propose here, needs to be administered in a controlled setting to be informative.

B: Jill Carroll was abducted in Iraq.

and

- A: At least 35 people were killed and 125 injured in three explosions targeting tourists in Egypt's Sinai desert region late Thursday, an Egyptian police source said.
  B: At least 30 people were killed in the blasts.

Some reasoning ability and background knowledge — in arithmetic, in geography — are necessary to get these questions correct. Nevertheless, it seems overly generous to rate a system as being on par with human intelligence on the basis of its ability to do well on a challenge of this difficulty. Certainly, this seems to be far below the difficulty level of what Turing was proposing.

What we propose in this paper is a variant of the RTE that we call the *Winograd Schema* (or WS) challenge. It requires subjects to answer binary questions, and appeals to world knowledge and default reasoning abilities, but without depending on an explicit notion of entailment.

4 The Winograd Schema Challenge

A WS is a small reading comprehension test involving a single binary question. Two examples will illustrate:

- The trophy doesn't fit in the brown suitcase because it's too big. What is 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

We take it that the correct answers here are obvious. In each of the questions, we have the following four features:

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 (but repeated from the sentence for clarity), 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.

We will explain the fourth feature in a moment. But note that like the RTE 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, like in the two examples above. (We will return to this point in the Incremental Progress section below.)

Perhaps the hardest item to justify even informally from the requirements in the previous section is that *thinking* is required to get a correct answer with high probability. Although verbal dodges are not possible like in the original Turing Test, how do we know that there is not some trick that a programmer could exploit, for example, the word order in the sentence or the choice of vocabulary, or some other subtle feature of English expressions? Might there not be some unintended bias in the way the questions are formulated that could help a program answer without any comprehension?

This is where the fourth requirement comes in. In 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:

- The trophy doesn't fit in the brown suitcase because it's too small. What is too small?

  Answer 0: the trophy
  Answer 1: the suitcase

(There is an extensive discussion of the spatial reasoning involved in these disambiguations in (Davis 2012).)

- 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. Contexts where "give" can appear are statistically quite similar to those where "receive" can appear, and yet the answer must change. This helps make the test *Google-proof:* having access to a large corpus of English text would likely not help much (assuming, that answers to the questions have not yet been posted on the Web, that is)! The claim is that doing better than guessing requires subjects to figure out what is going on: for example, a failure to fit is caused by one of the objects being too big and the other being too small, and they determine which is which.

big hope

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 (Winograd 1972), for whom we have named the schema:[2]

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?

[2]See also the discussion of this in (Pylyshyn 1984).

Answer 0: the town councillors
Answer 1: the angry demonstrators

It is wildly implausible that there would be statistical or other properties of the special word or its alternate that would allow us to flip from one answer to the other in this case. This was the whole point of Winograd's example! 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*. The fact that we are normally not *aware* of the thinking we are doing in figuring this out should not mislead us; using what we know is the only explanation that makes sense of our ability to answer here.

5 A library in standard format

In constructing a WS, it is critical to find a *pair* of questions that differ in one word and satisfy the four criteria above. In building a library of suitable questions, it is convenient therefore to assemble them in a format that lists both the special word and its alternate. Here is the first example above in this format:

The trophy doesn't fit in the brown suitcase because it's too ⟨⟩. What is too ⟨⟩?

Answer 0:	the trophy
Answer 1:	the suitcase
special:	big
alternate:	small

The ⟨⟩ in a WS is a placeholder for the special word or its alternate, given in the first and second rows of the table below the line. A WS includes both the question and the answer: Answer 0 (the first party in the sentence) is the correct answer when the special word replaces the ⟨⟩ and Answer 1 (the second party) is the correct answer when the alternate word is used.

While a WS involves a pair of questions that have opposite answers, 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:

- Paul tried to call George on the phone, but he wasn't ⟨⟩. Who wasn't ⟨⟩?

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

In putting together an actual test for a subject, we would want to choose randomly between the special word and its alternate. Since each WS contains the two questions and their answers, a random WS test can be constructed, administered, and graded in a fully automated way. An expert judge is not required to interpret the results.

6 What is obvious?

The most problematic aspect of this proposed challenge is coming up with a list of appropriate questions. Like the RTE, candidate questions will need to be tested empirically before they are used in a test. We want normally-abled adults whose first language is English to find the answers obvious. But what do we mean by "obvious"? There are two specific pitfalls that we need to avoid.

6.1 Pitfall 1

The first pitfall concerns questions whose answers are in a certain sense too obvious. These are questions where the choice between the two parties can be made without considering the relationship between them expressed by the sentence. 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 merely finding the permissible links between the answers and the special word (or its alternate). In linguistics terminology, the anaphoric reference can be resolved using selectional restrictions alone. Because selectional restrictions like this might be learned by sampling a large enough corpus (that is, by confirming that the word "pregnant" occurs much more often close to "women" than close to "pills"), we should avoid this sort of question.

Along similar lines, consider the following WS:

The racecar zoomed by the school bus because it was going so ⟨⟩. What was going so ⟨⟩?

Answer 0:	the racecar
Answer 1:	the school bus
special:	fast
alternate:	slow

In principle, both a racecar and a school bus can be going fast. However, the association between racecars and speed is much stronger, and again this can provide a strong hint about the answer to the question. So it is much better to alter the example to something like the following:

The delivery truck zoomed by the school bus because it was going so ⟨⟩. What was going ⟨⟩?

Answer 0:	the delivery truck
Answer 1:	the school bus
special:	fast
alternate:	slow

This pitfall can also be avoided by only using examples with randomly chosen proper names of people (like Joan/Susan or Paul/George, above) where there is no chance of connecting one of the names to the special word or its alternate.

6.2 Pitfall 2

The second and more troubling pitfall concerns questions whose answers are not obvious enough. Informally, a good question for a WS is one that an untrained subject (your Aunt Edna, say) can answer immediately.

But to say that an answer is obvious does not mean that the other answer has to be *logically inconsistent*. It is possible that in a bizarre town, the councillors are advocating violence and choose to deny a permit as a way of expressing this. It is also possible that angry demonstrators could nonetheless fear violence and that the councillors could use this as a pretext to deny them a permit. But these interpretations are farfetched and will not trouble your Aunt Edna.[3] So they will not cause us statistical difficulties except perhaps with language experts asked to treat the example as an object of professional interest.

To see what can go wrong with a WS, however, let us consider an example that is a "near-miss." We start with the following:

> Frank was jealous when Bill said that he was the winner of the competition. Who was the winner?
>
> Answer 0: Frank
> Answer 1: Bill

So far so good, with "jealous" as the special word and Bill as the clear winner. The difficulty is to find 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 (and similarly for "happy" or "overjoyed"). The sentence is too ambiguous to be useful. If we insist on using a WS along these lines, here is a better version:

> Frank felt ⟨ ⟩ when his longtime rival Bill revealed that he was the winner of the competition. Who was the winner?
>
> Answer 0: Frank
> Answer 1: Bill
>
> special: vindicated
> alternate: crushed

In this case, it is advisable to include the information that Bill was a longtime rival of Frank to make it more apparent that Frank was the winner.[4]

[3] Similarly, there is a farfetched reading where a small trophy would not "fit" in a big suitcase in the sense of fitting closely, the way a big shoe is not the right fit for a small foot.

[4] However, the vocabulary is perhaps too rich now.

7 Incremental Progress

In the end, what a subject will consider to be obvious will depend to a very large extent on what he or she knows. We can construct examples where very little needs to be known, like the trophy example, or this one:

> The man couldn't lift his son because he was so ⟨ ⟩. Who was ⟨ ⟩?
>
> Answer 0: the man
> Answer 1: his son
>
> special: weak
> alternate: heavy

At the other extreme, we have examples like the town councillor one proposed by Winograd. Unlike with the RTE, 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 background knowledge assumed. Consider, for example, this intermediate case:

> 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.

A major advantage of the WS challenge is that it allows for incremental progress. Like the RTE, it can be *staged*: we can have libraries of questions suitable for anyone who is at least ten-years old (like the trophy one), all the way up to questions that are more "university-level" (like the town councillor one). To get a feel for some of the possibilities, we include a number of additional examples in the Appendix at the end of the paper; a collection of more than 100 examples can be found at http://www.cs.nyu.edu/faculty/davise/papers/WS.html.

In addition, the schema can be grouped according to domain. Some examples involve reasoning about knowledge and communication; others involve temporal reasoning or physical reasoning. Researchers can choose to work on examples in a particular domain, and to take a test restricted to that domain.

To help ensure that researchers can make progress on the WS challenge at first, we propose to make publicly available well beforehand a list of *all the words* that will appear in a test. (Of course, we would include both the special words and their alternates, although only one of them will be selected at random when the test is administered.) For a test with 50 questions, which should be enough to rule out mere guessing, 500 words (give or take proper names) should be sufficient. A test with 50 questions should only take a person 25 minutes or so to complete.

8 Summary of a Winograd Schema

To summarize: A Winograd schema is a pair of sentences differing in only one or two words and containing an ambiguity that is resolved in opposite ways in the two sentences and that requires the use of world knowledge and reasoning for its resolution. It should satisfy the following constraints:

1. It should be easily disambiguated by the human reader. Ideally, this should be so easy that the reader does not even notice that there is an ambiguity; a "System 1" activity, in Kahneman's terminology (Kahneman 2011).
2. It should not be solvable by simple techniques such as selectional restrictions.
3. It should be Google-proof; that is, there should be no obvious statistical test over text corpora that will reliably disambiguate these correctly.

The proposed challenge would involve presenting a program claiming to intelligence with one sentence from every pair out of a hidden corpus of Winograd schemas. To pass the challenge, the program would have to achieve near human levels of success; presumably close to 100%, if constraint (1) above has been satisfied by the corpus designers.

The strengths of the challenge, as an alternative to the Turing test are that it is clear-cut, in that the answer to each schema is a binary choice; vivid, in that it is obvious to non-experts that a program that fails to get the right answers has serious gaps in its understanding; and difficult, in that it is far beyond the current state of the art.

9 Related Work

Alternatives to the Turing Test: (Cohen 2004), (Dennett 1998), (Ford and Hayes 1995), and (Whitby 1996) are among those who have argued against viewing the Turing Test as the ultimate test of artificial intelligence. Cohen has suggested several alternatives to the Turing Test, including a system capable of producing a five-page report on any arbitrary topic and systems capable of learning world knowledge through reading text. Unlike the WS Challenge, no definitive guidelines for success are given; passing the test would seem to rely on humna judgement. Dennett has observed that disambiguating Winograd-like sentences requires the sort of world knowledge and ability to reason we associate with intelligence, but has not expanded this observation into a proposal for an alternative to the Turing Test. A very different approach to testing intelligent systems, which uses principles of minimum length learning to develop a test applicable to *any* intelligence is presented in (Hernandez-Orallo and Dowe 2010).

Winograd Sentences: Sentences similar to Winograd Schema have been discussed by Hobbs (1979), Caramazza et al. (1977), Goikoetxea et al. (2008), and Rohde (2008). An example of Rohde's is:

> Mary scolded Sue. She kicked her.

She in the second sentence can refer to Mary (Mary scolded Sue, and and top of that Mary kicked Sue) or to Susan (Mary scolded Sue because Sue kicked Mary).

Carmazza et al. give examples of sentence pairs such as:

> Mary infuriated Jane because she had stolen a tennis racket.
> Mary scolded Jane because she had stolen a tennis racket.

which directly map onto the Winograd schema structure.

Much of this body of work focusses on exploring discourse coherence by classifying verbs by their role in a discourse context, since a verb's role can often give clues as to whether pronouns refer to the subject or an object of a previous sentence. For example, in the discourse fragment

> John infuriated Bill. He . . .

readers usually associate *He* with John; while in the discourse fragment

> John scolded Bill. He . . .

readers usually associate *He* with Bill. Readers expect the second sentence in the former fragment to *elaborate* on how John infuriated Bill, while they expect the second sentence in the latter fragement to *explain* why John scolded Bill; that is, what Bill had done to elicit the scolding. Goikoetxea, among others, discusses the implicit causal relational meaning in certain classes of verbs.

Hobbs is exemplary in noting the need for commonsense world knowledge to understand these sentences; however, the general focus of these researchers is on linguistic techniques. None have suggested using such pairs of sentences to test system comprehension and intelligence.

Datasets for testing understanding: Many datasets have been created to test systems' ability to reason. The best known of these are the RTE datasets.[5] As discussed in Section 2, the text and hypothesis pairs appear to focus on relatively shallow reasoning. The FRACAS dataset (Cooper et al. 1996) covers a greater range of entailment than the RTE datasets but is quite weak on anaphora reference, containing only a few such examples.

Variants of RTE: Choice of Plausible Alternatives (COPA) (Roemmele, Bejan, and Gordon 2011) is a proposed variant of the RTE challenge that focusses on choosing between two alternatives that ask about causes or consequences of a statement. Examples of COPA queries are:

> Premise: I knocked on my neighbor's door. *What happened as a result?*
> Alternative 1: My neighbor invited me in.
> Alternative 2: My neighbor left his house.

and

> Premise: The man fell unconscious. *What was the cause of this?*
> Alternative 1: The assailant struck the man in the head.
> Alternative 2: The assailant took the man's wallet.

Like the WS challenge, COPA emphasizes relatively deep reasoning. However, COPA's dataset includes problems that are less clear-cut than the WS schema. Determining success

[5] Recent RTE competitions have been organized by NIST. The site for the most recent 2011 competition, http://www.nist.gov/tac/2011/RTE/index.html, links to previous years' challenges. Papers describing systems, results, and samples of data are freely available; however, the data itself is generally available only to registered participants.

in COPA therefore requires a more standard NLP methodology: annotating examples, training on the annotation set, developing a human gold standard. In addition, COPA is narrower than the WS challenge: it focuses on causality, but makes no attempt to cover the broad range of human reasoning. It is not intended to supplant the Turing Test.

10 Discussion

10.1 Turing, Searle, and Behaviours

The claim of this paper in its strongest form might be this: with a very high probability, anything that answers correctly a series of these questions (without having extracted any hints from the text of this paper, of course) is thinking in the full-bodied sense we usually reserve for people.

To defend this claim, however, we would have to defend a philosophical position that Turing sought to avoid with his original Turing Test. So like Turing, it is best to make a weaker claim: with a very high probability, anything that answers correctly is engaging in behaviour that we would say shows thinking in people. Whether or not a subject that passes the test is really and truly thinking is the philosophical question that Turing sidesteps.

Not everyone agrees with Turing. Searle (2008) attempts to show with his well-known Chinese Room thought experiment that it is possible for people to get the observable behaviour right (in a way that would cover equally well the original Turing Test, an RTE test, and our WS challenge), but without having the associated mental attributes. However, in our opinion (Levesque 2009), his argument is vacuous: in particular, it is highly unlikely that a system without understanding that can accurately prescribe such complex behavior can be realized within the size of our universe.

On a related theme, Hawkins and Blakeslee (Hawkins and Blakeslee 2004) suggest that AI has focussed too closely on getting the behaviour right and that this has prevented it from seeing the importance of what happens *internally* even when there is no external behaviour. The result, they argue, is a research programme that is much too behavioristic. (Searle makes a similar point.) See also (Cohen 2004).

In our opinion, this is a misreading of Turing and of AI research. Observable intelligent behaviour is indeed the ultimate goal according to Turing, but things do not stop there. The goal immediately raises a fundamental question: what sorts of computational mechanisms can possibly account for the production of this behaviour? And this question may well be answered in a principled and scientific way by postulating and testing for a variety of internal schemes and architectures. For example, what are we to make of a person who quietly reads a book with no external behaviour other than eye motion and turning pages? There can be a considerable gap between the time a piece of background knowledge is first acquired and the time it is actually needed to condition behaviour, such as producing the answer to a WS.

10.2 Knowledge-based vs. Statistical Approaches

The computational architecture articulated by John McCarthy (McCarthy 1959) was perhaps the first to offer a plausible story about how to approach something like the WS challenge, introducing what is what is often called the *knowledge-based* approach (Brachman and Levesque 2004, Chap. 1): explicitly representing knowledge in a formal language, and providing procedures to reason with that knowledge. While this approach still faces tremendous scientific hurdles, we believe it remains the most likely path to success. That is, we believe that in order to pass the WS Challenge, a system will need to have commonsense knowledge about space, time, physical reasoning, emotions, social constructs, and a wide variety of other domains. Indeed, we hope that the WS Challenge will spur new research into representations of commonsense knowledge.

However, nothing in the WS challenge insists on this approach, and we would expect NLP researchers to try different approaches. Statistical approaches toward natural language processing (Manning and Schütze 1999) have become increasingly popular since the 1990s. Virtually all entrants to competitions like TREC (http://trec.nist.gov), and RTE have statistical components at their core; this is true even for natural language programs that emphasize the importance of knowledge representation and reasoning, such as the DARPA Machine Reading Program (Strassel et al. 2010), (Etzioni, Banko, and Cafarella 2006). The successes of the last several decades in such NLP tasks as text summarization and question-answering have been based on statistical NLP.

These successes have been on limited tasks and generally do not extend to the type of deep reasoning that we believe is required to solve the WS Challenge. But if statistical approaches over large corpora —- to gather commonsense knowledge or to learn patterns of pronoun referents — work better, so be it. The WS Challenge is agnostic about this matter. This agnosticism also means that we do not intend to provide training annotations.

10.3 Natural vs. Artificial Examples

The trend in natural-language processing challenges, such as RTE, TREC, and Machine Reading has been toward texts occurring naturally, such as newspaper articles and blog data. In contrast, the Winograd Schema set of examples is artificially constructed. However, we feel quite confident that the issues that arise in solving the Winograd schemas in our collection come up as well in interpreting naturally occurring text. Indeed, it is sometimes possible to find sentences in natural text that can easily be turned into Winograd schemas. Consider the following sentence from Jane Austen's *Emma*:

> Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses; and her place had been taken by an excellent woman as governess, who had fallen little short of a mother in affection.

This can be turned into the following WS schema:

> Emma's mother had died long ago, and her ⟨ ⟩ by an excellent woman as governess. Whose ⟨ ⟩ by the governess?

Answer 0:	Emma's mother
Answer 1:	Emma
special:	place had been taken
alternate:	education had been managed

Note also that disambiguating the second and third occurences of "her" in the original quotation, referring respectively to Emma and to Emma's mother, requires inference and world knowledge no less deep; however, these do not seem to be easily transformable into Winograd schemas.

The difficulty is that there are certain conventions in text in general, and probably more specific conventions in the works of particular authors, which can be exploited by a system that attempts at no comprehension, but merely uses statistical knowledge. For example, Hobbs (1979) cites studies that show that in naturally-occurring text, an ambiguous pronoun more often refers to the subject of the preceding sentence than the object. More exact figures can doubtless be determined from studies of individual authors.

While we are not opposed to the use of statistical methods, we do not believe that systems that use statistics alone, in the absence of world knowledge and any method that simulates reasoning, are conforming to the spirit of the test. Artificially constructing examples allows the test designer to prevent test takers from using knowledge-free statistical methods.

The major disadvantage of using a hand-crafted test set is that it can be expensive to construct large test sets. This might be a problem if we were intending to construct large sets at very frequent intervals —e.g., if we were envisioning holding a yearly competition with large training and test sets. But since we don't envision doing that, and since the labor involved in constructing a small dataset of around 100 examples is on the order of one or two weeks of work, we do not consider this to be much of an issue.

10.4 Conclusion

Like Turing, we believe that getting the behaviour right is the primary concern in developing an artificially intelligent system. We further agree that English comprehension in the broadest sense is an excellent indicator of intelligent behaviour. Where we have a slight disagreement with Turing is whether a free-form conversation in English is the right vehicle. Our WS challenge does not allow a subject to hide behind a smokescreen of verbal tricks, playfulness, or canned responses. Assuming a subject is willing to take a WS test at all, much will be learned quite unambiguously about the subject in a few minutes. What we have proposed here is certainly less demanding than an intelligent conversation about sonnets (say), as imagined by Turing; it does, however, offer a test challenge that is less subject to abuse.

11 Acknowledgements:

An earlier version of this paper, with Hector Levesque as sole author, was presented at Commonsense-2011. We thank Ray Jackendoff, Mitch Marcus, and the anonymous reviewers of this paper for their helpful suggestions and comments.

Appendix A: Corpus of Winograd schemas

This appendix gives some examples of the more than 100 additional Winograd schemas available at http://www.cs.nyu.edu/faculty/davise/papers/WS.html. [6] In the interests of space, we have adopted a more compact format. In some cases where we were concerned that the schema might not be Google-proof, we have done some experiments with searches using Google's count of result pages. These counts, however, are notoriously unreliable (Lapata and Keller 2005), so these "experiments" should be taken with several grains of salt.

1. John couldn't see the stage with Billy in front of him because he is so [short/tall]. Who is so [short/tall]?
   Answers: John/Billy.
2. Tom threw his schoolbag down to Ray after he reached the [top/bottom] of the stairs. Who reached the [top/bottom] of the stairs?
   Answers: Tom/Ray.
3. 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.
4. The sculpture rolled off the shelf because it wasn't [anchored/level]. What wasn't [anchored/level]?
   Answers: The sculpture/the shelf.
5. 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.
6. 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
7. 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.
8. 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.
9. Jim [yelled at/comforted] Kevin because he was so upset. Who was upset?
   Answers: Jim/Kevin.
10. 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.
11. Pete envies Martin [because/although] he is very successful. Who is very successful?
    Answers: Martin/Pete.

[6] Thanks to Pat Levesque and reviewers for help with the first several examples and to Stavros Vassos for general discussion.

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

    Answers: the table/the cloth.

13. Sid explained his theory to Mark but he couldn't [convince/understand] him. Who did not [convince/understand] whom?

    Answers: Sid did not convince Mark/Mark did not understand Sid.

14. Susan knew that Ann's son had been in a car accident, [so/because] she told her about it. Who told the other about the accident?

    Answers: Susan/Ann.

15. The drain is clogged with hair. It has to be [cleaned/removed]. What has to be [cleaned/removed]?

    Answers: The drain/the hair.

16. My meeting started at 4:00 and I needed to catch the train at 4:30, so there wasn't much time. Luckily, it was [short/delayed], so it worked out. What was [short/delayed]?

    Answers: The meeting/the train.

17. There is a pillar between me and the stage, and I can't [see/see around] it. What can't I [see/see around]?

    Answers: The stage/the pillar.

18. Ann asked Mary what time the library closes, [but/because] she had forgotten. Who had forgotten?

    Answers: Mary/Ann.

19. Bob paid for Charlie's college education, but now Charlie acts as though it never happened. He is very [hurt/ungrateful]. Who is [hurt/ungrateful]?

    Answers: Bob/Charley

20. At the party, Amy and her friends were [chatting/barking]; her mother was frantically trying to make them stop. It was very strange behavior. Who was behaving strangely?

    Answers: Amy's mother/Amy and her friends.

21. The dog chased the cat, which ran up a tree. It waited at the [top/bottom] Which waited at the [top/bottom]?

    Answers: The cat/the dog.

22. Sam and Amy are passionately in love, but Amy's parents are unhappy about it, because they are [snobs/fifteen]. Who are [snobs/fifteen]?

    Answers: Amy's parents/Sam and Amy.

23. Mark told Pete many lies about himself, which Pete included in his book. He should have been more [truthful/skeptical]. Who should have been more [truthful/skeptical]?

    Answers: Mark/Pete.

24. Since it was raining, I carried the newspaper [over/in] my backpack to keep it dry. What was I trying to keep dry?

    Answers: The backpack/the newspaper.

25. Jane knocked on Susan's door, but she didn't [answer/get an answer]. Who didn't [answer/get an answer]?

    Answers: Susan/Jane.

26. Sam tried to paint a picture of shepherds with sheep, but they ended up looking more like [dogs/golfers]. What looked like [dogs/golfers]?

    Answer: The sheep/the shepherds.

27. Thomson visited Cooper's grave in 1765. At that date he had been [dead/travelling] for five years. Who had been [dead/travelling] for five years?

    Answers: Cooper/Thomson

28. Tom's daughter Eva is engaged to Dr. Stewart, who is his partner. The two [doctors/lovers] have known one another for ten years. What two people have known one another for ten years?

    Answers: Tom and Dr. Stewart / Eva and Dr. Stewart.

29. The actress used to be named Terpsichore, but she changed it to Tina a few years ago, because she figured it was [easier/too hard] to pronounce. Which name was [easier/too hard] to pronounce?

    Answers: Tina/Terpsichore.

30. Sara borrowed the book from the library because she needs it for an article she is working on. She [reads/writes] when she gets home from work. What does Sara [read/write] when she gets home from work/

    Answers: The book/the article.

31. Fred is the only man still alive who remembers my great-grandfather. He [is/was] a remarkable man. Who [is/was] a remarkable man?

    Answers: Fred/my great-grandfather.

32. Fred is the only man alive who still remembers my father as an infant. When Fred first saw my father, he was twelve [years/months] old. Who was twelve [years/months] old?

    Answers: Fred/my father.

33. There are too many deer in the park, so the park service brought in a small pack of wolves. The population should [increase/decrease] over the next few years. Which population will [increase/decrease]?

    Answers: The wolves/the deer.

34. Archaeologists have concluded that humans lived in Laputa 20,000 years ago. They hunted for [deer/evidence] on the river banks. Who hunted for [deer/evidence]?

    Answers: The prehistoric humans/the archaeologists.

35. The scientists are studying three species of fish that have recently been found living in the Indian Ocean. They [appeared/began] two years ago. Who or what [appeared/began] two years ago?

    Answers: The fish/the scientists.

36. The journalists interviewed the stars of the new movie. They were very [cooperative/persistent], so the interview lasted for a long time. Who was [cooperative/persistent]?

    Answers: The stars/the journalists.

37. I couldn't find a spoon, so I tried using a pen to stir my coffee. But that turned out to be a bad idea, because it got full of [ink/coffee]. What got full of [ink/coffee]?

    Answers: The coffee/the pen.

Comment: The statistical associations give the backward answer here: "ink" is associated with "pen" and "coffee" is associated with "coffee". Of course, a contestant could use a backward rule here: Since the challenge designers have excluded examples where statistics give the right answer, if you find a statistical relation, guess that the answer runs opposite to it. But that seems very risky.

References

Bobrow, D.; Condoravdi, C.; Crouch, R.; de Paiva, V.; Karttunen, L.; King, T.; Mairn, B.; Price, L.; and Zaenen, A. 2007. Precision-focussed Textual Inference. In *Proc. Workshop on Textual Entailment and Paraphrasing*.

Brachman, R., and Levesque, H. 2004. *Knowledge Representation and Reasoning*. Morgan Kaufman.

Christian, B. 2011. Mind vs. Machine. *Atlantic Monthly*. March 2011.

Cohen, P. 2004. If not the Turing Test, Then What? In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*. Menlo Park, Calif.: AAAI Press.

Cooper, R.; Crouch, D.; Eijckl, J. V.; Fox, C.; Genabith, J. V.; Japars, J.; Kamp, H.; Milward, D.; Pinkal, M.; Poesio, M.; and Pulman, S. 1996. A Framework for Computational Semantics (FraCaS). Technical report, The FraCaS Consortium.

Dagan, I.; Glicksman, O.; and Magnini, B. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: LNAI 3944*. Springer Verlag.

Davis, E. 2012. Qualitative Spatial Reasoning in Interpreting Text and Narrative. *Spatial Cognition and Computation*. Forthcoming.

Dennett, D. 1998. Can Machines Think? In Mather, G.; Verstraten, F.; and Anstis, S., eds., *The Motion Aftereffect*. MIT Press.

Etzioni, O.; Banko, M.; and Cafarella, M. 2006. Machine Reading. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*. Menlo Park, Calif.: AAAI Press.

Ford, K., and Hayes, P. 1995. Turing Test Considered Harmful. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 972–977. San Mateo, Calif.: Morgan Kaufmann.

Hanks, S., and McDermott, D. 1987. Nonmonotonic Logic and Temporal Projection. *Artificial Intelligence* 33(3):379–412.

Hawkins, J., and Blakeslee, S. 2004. *On Intelligence*. New York: Times Books.

Hernandez-Orallo, J., and Dowe, D. L. 2010. Measuring Universal Intelligence: Toward an Anytime Intelligence Test. *Artificial Intelligence* 174(18):1508–1539.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus, and Giroux.

Lapata, M., and Keller, F. 2005. Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing* 2(1).

Levesque, H. 2009. Is it Enough to get the Behaviour Right? In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence*. San Mateo, Calif.: Morgan Kaufmann.

Majumdar, D., and Bhattacharyya, P. 2010. Lexical Based Text Entailment System for Main Task of RTE6. In *Proceedings, Text Analysis Conference, NIST*.

Manning, C., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

McCarthy, J. 1959. Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*. London: Her Majesty's Stationery Office.

Pylyshyn, Z. 1984. *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Mass.: MIT Press.

Roemmele, M.; Bejan, C.; and Gordon, A. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proceedings, International Symposium on Logical Formalizations of Commonsense Reasoning*.

Rus, V.; McCarthy, P.; McNamara, D.; and Graesser, A. 2007. A Study of Textual Entailment. *International Journal of Artificial Intelligence Tools* 17.

Shieber, S. 1994. Lessons from a Restricted Turing Test. *Communications of the ACM* 37(6):70–78.

Strassel, S.; Adams, D.; Goldberg, H.; Herr, J.; Keesing, R.; Oblinger, D.; Simpson, H.; Schrag, R.; and Wright, J. 2010. The DARPA Machine Reading Program - Encouraging Linguistic and Reasoning Research with a Series of Reading Tasks. In *International Conference on Language Resources and Evaluation (LREC)*.

von Ahn, L.; Blum, M.; Hopper, N.; and Langford, J. 2003. CAPTCHA: Using Hard AI Problems for Security. In *Eurocrypt-2003*, 294–311.

Weizenbaum, J. 1966. ELIZA — A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM* 9(1):36–45.

Whitby, B. 1996. Why the Turing Test is AI's Biggest Blind Alley. In Millican, P., and Clark, A., eds., *Machine and Thought*. Oxford University Press.

Winograd, T. 1972. *Understanding Natural Language*. New York: Academic Press.