What can go wrong with the Winograd Schema Challenge, and how not to fix it

Ernest Davis
Department of Computer Science
New York University
davise@cs.nyu.edu

February 1, 2015

The Best Case

The ideal outcome of the Winograd Schema Challenge (WSC), in the minds of its creators, or at least in my mind, would be something along the following lines: The announcement of the WSC will be the bellwether of a rapidly growing movement, in which the AI research community, seeing past the false promise and limitations of corpus-based learning, turns its attention back to the foundational problems of commonsense reasoning and its relation to natural language understanding; this time making steady progress. CYC will be cleaned up; NELL and KnowItAll will be made deeper; ConceptNet will be made more discriminating; I will finish my book on ontologies of matter; and so on; and they will all be integrated into a system that will pretty much sew up KR and its relation to NLP. At that point, the Winograd Schema Challenge will hardly matter any more; there will be programs available that can solve it without breaking a sweat.

In this note, I want to discuss ways in which the future of the WSC may fall short of this beatific vision. I will argue that it may be a total flop, in a number of ways, though I doubt it; but if it is, there is no point in trying to patch it or change the rules to avoid that; all that can be done is to scrap it.

The Worst Case

First, let me make the obvious point that a program that only passed the WSC — that is, a program that could get the correct answer to any WS, but could do nothing else, and whose development made no other contribution to AI — would be both entirely useless and entirely uninteresting. This is in contrast to, say, a translation program, or a self-driving car, which would be useful things to have if we got them from the Martians, got nothing else, and had no idea how they worked. It is also in contrast to Deep Blue and Watson, which were at least impressive, even if not very useful.

Therefore, the worst possible scenario would be one in which someone managed to pass the WSC using a technique that worked *only* to disambiguate sentences that are one half of a WS, and that made no contribution at all, either to commonsense reasoning or coreference resolution, either theoretical or practical. That is, it relies on the fact that the sentences presented are half

of a WS (otherwise, presumably, it would be of some value in pronoun resolution — we will return to this). If that were the outcome of the WS Challenge, then we would have accomplished absolutely nothing by announcing it except to sow confusion, hand out undeserved bragging rights, waste time, and waste the money of the Nuance Corporation.

This seems quite implausible, for the following reason: Recall that a contestant in the WSC sees only *one* of the two sentences of the schema. Therefore, there should be very little difference between solving a Winograd schema, and solving pronoun resolution generally, except that the sentences in the WSC have been chosen so that "easy" resolution techniques such as frequency based techniques and selectional restrictions don't work.

One might ask then, why bother with Winograd schemas at all; why not simply use sentences where the resolution seems to be difficult? Why adopt such the strange, highly limiting constraint that we need two sentences, when we are only going to use one of them? The answer is that the purpose of the Winograd Schema format has nothing to do with the contestant; what the contestant is seeing *is* just a sequence of difficult resolution problems. The point of the WS format is as a crutch to the challenge developer; it gives him a framework in which he/she can be fairly sure that the easy techniques won't work. We can be sure, for example, that the recency heuristic has zero value in the WSC, because in each pair there is at least one where the heuristic is not satisfied[1] In the many Winograd schemas that involve two people as antecedents, simple selectional restrictions have demonstrably zero value, since the two antecedents have exactly the same unary features.

However, it is not entirely impossible that a program might be able to do resolution on sentences that it realized were half of a WS, but otherwise had no edge on doing pronoun resolution. Let me sketch three ways this might occur.

The first scenario is that sentences that are half of a WS have, because of the constraints, some kind of detectable feature that gives away the answer, a feature that is of no value for other sentences with ambiguous pronouns. Now, there certainly are distinctive statistical regularities among the WS's we have collected; e.g. in a disporportionate fraction the "special" word is the last word, or nearly the last word. One can imagine, let us say, a program that used that kind of fact, plus the fact that this *is* half of a WS, to figure out what the other half must be; it then has a comparative problem to solve, which could well often be easier than resolving the referent in a single sentence. This scenario seems wildly unlikely however. It is almost impossible to believe that it would be easier to build a program that can find the other half of a WS than to build one that can solve a WS; and more generally, (and more vaguely) the distinctive statistical patterns that one sees in WS's do not seem to shed much light on how the ambiguity should be resolved; not least, because these patterns generally turn up in both sentences (e.g. the fact that the special word is the last word.)

The second scenario is, unfortunately, much less implausible; it is simply the possibility that there *are* only finitely many Winograd schemas, up to trivial variation. Given the multiple constraint that a Winograd schema has to satisfy, I don't see any way to rule out this possibility. If this is true, then it would be possible to solve the WSC, simply by enumerating all the possibilities (if he missed a few it wouldn't matter, since they would be unlikely to be in the test collection anyway) and hand coding up a separate technique for each special case. I don't see any way to estimate the probability of this.

One significant consolation here is that, though all the other consequences of a flop that I enumerated above would apply, and we would certainly come out with egg on our shoulder, it is hard to see that the winner would have garnered any significant bragging rights. It would be

[1]Most of the time, there is one where it is satisfied and one where it is not; but there could be examples where the recency heuristic indicated an antecedent that was wrong for both sentences.

clear enough that, though he had outsmarted Hector, Leora, and me, he had not actually made any contribution to NLP or to AI. Note that we do require a winning contestant to publish an account of how the program works. So this should not lead to a general impression that some great leap forward in AI had been made. It might lead to grumblings from Geoff Hinton that we had moved the goalposts; but I don't think we need to take those very seriously.

Another thought is that, just as I can't have any confidence that it is false, I don't see how the next guy could have any confidence that it is true in general, or that at a certain point he had found all or most of the possible Winograd schemas. So it seems unlikely that anyone would actually set out to do this, particularly since each of the "special cases" that I waved at airily above would be a fair amount of work to deal with, and since, when he was all done, he would not have accomplished anything at all worth doing anyway. So, in practical terms, I think this is unlikely to be an issue.

A third possibility is techniques that rely on stylistic quirks of the particular individuals who are making up the Winograd schemas. Maybe all the schemas that I make up have the feature that the special word and the alternate word are from opposite ends of the alphabet, and the recency heurstic is observed for the sentence in the pair where the low alphabetical word is chosen and violated for the sentence where the high alphabetical word is chosen. But it seems to me unlikely that enough statistical leverage on the problem can be gotten this way because the the training set of Winograd schemas that I've made up and published is necessarily quite small. Also, again, a program that relied on this would clearly be using a cheap trick, and so would hardly garner much favorable publicity.

There could also certainly be other methods that I haven't thought of that will solve specifically halves of WS's but not hard pronoun resolution problems generally. But it seems improbable to me that any such technique could get enough statistical leverage to get a high score on the WSC.

For this reason, it is important that we emphasize publically that scientists should be studying pronoun resolution, and particularly the use of pragmatic world knowledge in pronoun resolution; scientists should not be studying pronoun resolution in Winograd schemas specifically. In the first place, we do not believe that studying Winograd schemas specifically will lead to a solution of the WSC; and second if it does, it would be an entirely sterile solution, and a waste of time. I would really rather not be reviewing a yearly collection of papers entitled "Solving the Winograd Schema Challenge", at whatever cost to my citation count.

In Between

If a program is to achieve a high score on the WSC, and it uses methods that do not particularly apply to WS halves as opposed to other hard pronoun resolution problem, then it is a technique that is extremely successful on hard resolution problems generally. Now, that *would* be an important accomplishment in natural language processing, whatever technique is used, and I would not have any reservations about awarding prizes, recognition, etc. to the program that accomplished it, even if it uses entirely statistical/Big Data methods with no hint of knowledge representation or commonsense reasoning. I would be disappointed, of course.

However, I think that such a scenario is extremely unlikely. If statistical techniques suffice for reference disambiguation, then I would be ready to believe that they suffice for intelligence generally. It's in this sense that I think the WSC is a reasonable substitute for the Turing test; not, obviously that, a program that can do pronoun resolution has all the intelligence of a person; but that if we can crack pronoun resolution, we can crack a lot of what is considered to require deep knowledge and high-level reasoning.

Inherent limits on the WSC

Of course, there are all kinds of intelligence, and even many forms of commonsense reasoning, that are hard or impossible to test using reference resolution examples. One cannot simply take a fact of commonsense knowledge, a commonsense inference, or even a domain of commonsense knowledge, and expect to be able to design a plausible sentence whose disambiguation relies on that fact, let alone to design a full Winograd Schema. Moreover, the knowledge that is used is in many ways shallow; in terms of spatial reasoning, for example, it is easy to design WS's that rely on an understanding of big vs. small, high vs. low and so on; I have found it impossible to design one that depends on a sophisticated understanding of shape. So it is quite possible that a commonsense knowledge base sufficient for essentially all pronoun resolution tasks would use a much more anemic spatial theory than would be needed for, say, video understanding. That is inherent to the task of pronoun resolution; so relying on this fact in building a program that does the WS Challenge cannot be considered gaming.

Ideas for combatting gaming

The suggestion has been made that, since there is some possibility that the WSC can be "gamed", the format should be revised to include some more stringent test that the program actually understands the questions and is using world knowledge in its disambiguation. For instance it has been proposed that a program has to provide a trace of its reasoning that is inspected by judges; or that it has to explain its reasoning; or that it has to answer follow-up questions about the sentence.

I think the drawbacks of all of these proposals are much greater than the problems with the WSC that they solve. (Note that if such proposals are implemented, the probability of the drawbacks is 1, whereas the probability that the WSC in its original form is flawed is less than 1.)

First: One of the great features of the WSC is the simplicity of evaluation: It is a score on a sequence of questions with two answers and an unequivocal gold standard. No partial credit, no subjective evaluation. That great advantage is thrown away.

Second: It makes the test much harder for the human subjects that are being used as a gold standard. "It is easier to know it than to explain why I know it" (*A Study in Scarlet*). Of course, it's impossible for people to provide a trace of their own thinking. As for giving an explanation, consider schema #19: "The sack of potatoes had been placed above the bag of flour, so it had to be moved first. What had to be moved first?" "Answer: The sack of potatoes." "How do you know it wasn't the bag of flour", "Well because the sack of potatoes was on top, so of course it would have to be moved first. It wouldn't make any sense the other way." I don't think you will get anything better out of a lot of subjects; and (a) this would be frustrating for subjects; (b) this level of "explanation" is probably a lot easier to game than getting the answer right. As for follow-up questions, it depends a lot on the question. I don't actually see any good follow-ups for this one.

Third: It creates a great deal more work on the challenge designers, since they would have to provide guidelines and examples for proper traces / explanations / follow-up questions with answers.

Fourth: Traces in particular run against Turing's idea, which I think we want to stick to, that intelligence is judged by what tasks a program can carry out, not by how it is carrying them out.

Fifth: Suppose that a program passes the WSC but fails the extension. Then we get into a

big debate about whether we've made the right judgement etc.

So my bottom line is that any of the extensions to the WSC hugely complicates things for the designers, for the contestants, for the judges, and for the human subjects, and weakens the impact of the challenge for the public, for very little gain.

The following, though, I think might be worth considering. We change the "Winograd Schema challenge" to the "hard pronoun resolution challenge" (so to speak: we can keep the actual name the same). We reserve the right to mix in sentences that raise difficult problems of pronoun resolution together with the halves of Winograd schemas. In practice we use a roughly even quantities, and we take the non WSs out of some natural source. That way, challengers that are working off some cheap trick based on being a half of a Winograd schema should not score higher than 75%, and challengers that are depending on features that we are missing in the natural text but that the WS format corrects for should also not score higher than 75%. So we would want to set the bar at 85% or so. If they pass this test, then they've presumably got a real good working solution to hard pronoun resolution, and I think they deserve a prize.

This will also have some ancillary advantages. 1) It will force contestants to work on pronoun resolution generally and not to focus on Winograd schemas specifically. 2) It makes it easier, either for us or, preferably, for someone else, to assemble a large training set. 3) It partially counters objections that this is not a natural task.