On the Winograd Schema: Situating Language Understanding in the Data-Information-Knowledge Continuum

Walid S. Saba

Astound.ai
111 Independence Drive
Menlo Park, CA 94025 USA
walid@astound.ai

Abstract
The Winograd Schema (WS) challenge has been proposed as an alternative to the Turing Test as a test for machine intelligence. In this paper we ‘situate’ the WS challenge in the data-information-knowledge continuum, suggesting in the process what a good WS is. Subsequently, we will argue that the WS is but a special case of a more general phenomenon in language understanding, namely the phenomenon of the ‘missing text’. In particular, we will argue that what we usually call thinking in the process of language understanding almost always involves discovering some missing text - text is rarely explicitly stated but is implicitly assumed as shared background knowledge. As such, we suggest extending the WS challenge to include other linguistic phenomena that also involve discovering the ‘missing text’, such as tests metonymy, quantifier scope, lexical disambiguation, and copredication, to name a few.

Introduction
Consider the sentence in (1):

(1) Dave told everyone in school that he wants to be a guitarist, because he thinks it is a great sounding instrument.

Short of accessing knowledge structures that somehow encode the commonsense facts that a-guitarist-plays-a-guitar and that a-guitar-is-a-musical-instrument, much like a 4-year old would do in understanding what (1) says, quantitative (i.e., statistical and data-driven machine learning) methods would, with a high degree of certainty, erroneously resolve “it” in (1) since the correct referent is not even in the data, but is an object that is implicitly assumed by the semantic and cognitive content of the text. Undoubtedly it is this kind of thinking that Alan Turing had in mind when he posed the question “Can Machines Think?” (Turing, 1950), suggesting further that a machine that comprehends and communicates in ordinary spoken language, much like humans do, must be a thinking machine.

While it was certainly ingenious of Turing to recognize that competency in comprehending ordinary spoken language is the best test for thinking machines, the imitation game (the “Turing Test”) that he proposed has several shortcomings. As recently suggested by Levesque et al. (2012), the Turing Test left room for the possibility of some systems to pass the test, not because any thinking is going on, but by trickery and deception. As Levesque points out, systems that have participated in the Loebner competition (Shieber 1994) usually use deception and trickery by throwing in “elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts,” while avoiding clear answers to questions that a 5-year old would be very comfortable in correctly answering. In addressing these shortcomings, Levesque et al suggested what they termed the Winograd Schema (WS) challenge, illustrated by the following example:

1 Although this is not the subject of the current discussion, we however would like to unequivocally concur that language, that infinite object that is tightly related to our capacity to have an infinite number of thoughts, is the ultimate test for thinking machines. Thus, and while several accomplishments in computation are usually attributed to AI, most of these tasks deal with finding a near optimal solution from a finite set of possibilities and are hardly performing what we might call thinking. For example, and although the search space is very large (and would therefore require some intelligent heuristics) playing chess is, ultimately, a matter of scoring more paths than the opponent can, thus making the probability of winning in the long run certain. The same can also be said of pattern (sound and image) recognition systems that, essentially, find regularities in data. True human-level scene analysis, beyond lower level image recognition that the most primitive of species can perform, would in the end also require reasoning that is somewhat similar to that required in language understanding. In this regard, we also believe the recently proposed visual Turing Test for vision systems (Geman et al., 2014) is a step in the right direction.

2 The example in (2) was originally discussed by Terry Winograd (1972), after whom the challenge was named.
The city councilmen refused the demonstrators a permit because they

a. feared violence.
b. advocated violence.

The question posed against this sentence would be: what does “they” refer to in the case of (2a), and what does it refer to in (2b)? The answer seems so obvious to humans that reason using commonsense (background) knowledge (that, for example, demonstrators are more prone to advocate violence than the governing body of a certain city, while the latter are more likely to fear the violence) and thus a machine that correctly resolves such references would be performing what we might call thinking. Levesque points out however that care should be taken in the design of such queries so as to avoid the pitfalls of the original Turing Test, namely that a program should not be able to pass the test by performing simple syntactic level and pattern matching computations. For example, simple word co-occurrence data obtained from a corpus analysis might be all that is needed to make the correct guess in (3), without doing anything we might call thinking, while the same cannot be done in (4).

The women stopped taking the pills because they were

a. pregnant.
b. carcinogenic.

The trophy would not fit into the brown suitcase because it was too

a. small.
b. big.

Levesque calls the query in (4) “Google-proof”, since having access to a large corpus would not help here as the frequencies of contexts where “small” and “big” appear should in principal be the same. This is not the same in (3), however, where a purely quantitative system would pass many queries based on simple co-occurrence data as the likelihood of “carcinogenic” co-occurring with “pills” should be much higher than its co-occurrence with “women” (and similarly for the other combination).

Another important point Levesque makes in proposing the WS challenge is avoiding posing queries that are either too obvious, or too difficult. The latter could happen, for example, if the questions posed required knowledge of a special vocabulary – a vocabulary that only specialized domain experts might know. Essentially, good WS sentences should be ones that a 5-year old would be able to effortlessly answer – or, as Levesque puts it, “a good question for a WS is one that an untrained subject (your Aunt Edna, say) can answer immediately”. This is a crucial issue and it is central to the design of a good set of WS sentences. The obvious question then is what kinds of sentences are appropriate sentences to the WS challenge. In the next section we suggest how this answer can be systematically answered.

---

**Figure 1.** The Winograd Schema situated in the data-information-knowledge continuum
The references in (6a) and (6b) can be easily resolved using information that is readily available in the data itself - in particular, the references in (6) can be resolved by ensuring gender (male/female/neutral) and number (singular/plural) agreement.

The Semantic/Information Level

In this level the information required is not readily available in the data itself, but is one step away, so to speak, as it is in attributes and properties of the lexical items in the data. Here is an example illustrating this situation:

(7) Our graduate students published 20 papers this year and, apparently, few of them
   a. also authored books
   b. appeared in top journals

Here type information (often referred to in the literature as ‘selectional restrictions’) is sufficient to resolve the anaphor ‘them’: students, not papers, author books; and papers, not students, appear in journals. This information, while not readily available in the text itself, is only one step away as such information can be represented as properties/attributes of the lexical items in the text: AUTHOR(person, publication) and PUBLISH(paper, journal), etc. While such information is not readily available for a quantitative (statistical/machine learning) system, these systems can make reasonable guesses (that cannot, in any case, scale into a workable solution) using probabilities of word co-occurrences computed over a very large corpus.

The Pragmatic/Knowledge Level

This is the level at which ‘good’ WS sentences are situated. Sentences at this level are those where the reference in question can, in theory, be resolved by either of the two noun phrases, where the ‘most appropriate’ referent is usually the one that is more probable than the other (or, the one that makes the final scenario being described more compatible with our commonsense understanding of the world we talk about in ordinary spoken language). Care should therefore be exercised in not choosing WS sentences where the likelihood of both referents are near equal (cases where the WS is too difficult), or where the likelihood of one is clearly much higher than the other (cases where the WS is too easy). Below are some examples, in order of difficulty, where ideal WS sentences are those that are in the middle.

Level 3.1 At this level both referents are in theory possible, but the likelihood of one is more plausible - that is, the likelihood of one referent is much higher than the
likelihood of the other. The sentences in (8) are typical examples.

(8) John told Bill he will be coming to the meeting.
While, in theory, both John and Bill are valid (and possible) referents for 'he', John is a lot more plausible, since it is more likely for someone to tell another a fact about themselves then to claim knowledge of someone else’s intentions. While such sentences could be good test cases in the WS challenge, we consider these examples to be the ‘easy’ test cases.

Level 3.2 Again, at this level both referents are also possible in theory. While the likelihood of one of the referent being the correct one is higher than the other choice, it is not much higher, and the correct referent is chosen because it is the one that is slightly more consistent with our commonsense understanding of the world. The sentences (9) through (11) are typical examples of this situation.

(9) Professor Carnap told John that he should soon be done with reviewing his thesis.

(10) John could not lift his son because he was too
   a. heavy
   b. weak

(11) The trophy would not fit into the brown suitcase because it was too
   a. big
   b. small

Again, as in all situations within level 3, the references in (9) through (11) can, in theory, be assigned to either of the possible referents in the sentence. However, in situations like the ones above the likelihood of one of the referents is usually slightly more consistent than the other with our commonsense knowledge of the world. These situations are the most interesting for a good set of WS sentences as they satisfy three important criteria: (i) these sentences cannot be handled at the data level by simple syntactic information; (ii) these examples are “Google proof” - that is, they cannot be resolved by relying on prior probabilities obtained by training on a large corpus; and (iii) in these sentences both referents are quasi-possible, yet one is more consistent with our commonsense knowledge of the world we talk about, and as such, a machine that can perform such inferences could be simulating what we might call thinking.

Level 3.3 Sentences at this level are similar to the ones in level 3.1 and 3.2 in that both referents are also possible in theory, and the likelihood of one of the referent being the correct one is slightly higher than the other choice. However, the difference in the degree of likelihood of either referent gradually becomes very small in such examples, to the point where it could become questionable as to why one made one choice over the other. Such “difficult” examples can perhaps be excluded from the sentences posed in a WS challenge. The sentences (12) and (13) are typical examples of this situation.

(12) The town councilors refused to give the angry demonstrators a permit because they
   a. feared violence
   b. advocated violence

(13) A young teenager fired several shots at a policeman. Eyewitnesses say he immediately
   a. fell down
   b. fled away

Although it is not very likely, one can still imagine (12) describing a scenario where some (ill-intentioned, or anarchist-dominated) city council refusing to give a permit for a (peaceful) demonstration, and specifically to incite violence. The sentence in (13) is even more of a typical example in this level, where both referents starts to be almost as likely: one can easily imagine a young teenager falling down after shooting a policemen with a very large machine gun, or a policemen fleeing away after being shot, perhaps to escape further injuries.

While sentences in levels 3.1, 3.2 and 3.3 are all examples of sentences where both referents are in theory possible, sentences in level 3.3 are examples of sentences where the resolution of the reference becomes too difficult, in contrast to the ones in level 3.1 where the resolution of the reference is perhaps too easy. It is thus sentences in level 3.2 that are good WS sentences. In particular, it is sentences where both referents are in principal possible, yet humans (like our 5-year old, or Aunt Edna!) can easily make the correct choice because the likelihood of one of the referents is, overall, more consistent with our commonsense understanding of the world.

---

3 By ‘likelihood’ we mean something like “the commonsense likelihood”. Thus “the likelihood of referent R being the correct one” essentially means the “the likelihood of referent R being selected is more consistent with our commonsense understanding of the world”.

4 Sentences in level 3, whether are considered ‘too easy’, ‘too difficult’ or ‘just right’ could all be part of the WS challenge, and a weighted scoring can take their level of difficulty into consideration.

5 Ironically, and while purely statistical (or other quantitative) models would not be able to pass good WS sentences, thinking seems to be happening in sentences where humans seem to be good at choosing the meaning that is, statistically, the most consistent with our commonsense understanding of the world.
The Phenomenon of the ‘Missing Text’: is the Winograd Schema just a Special Case?

Having discussed the Winograd Schema (WS) in some detail, suggesting in the process where good WS sentences are situated, we would like to suggest here that WS sentences are in fact special cases of a more general phenomena in natural language understanding that a good test for intelligence must consider.

The sentences in level 3.2 (which is a subset of level 3 in figure 1) are good WS sentences specifically because these are typical examples where humans perform commonsense reasoning accessing background knowledge that is not explicitly stated in the text. As Levesque (2012), noted:

“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 it is precisely bringing this background knowledge to bear that we informally call thinking.” (Emphasis added)

We wholeheartedly agree: accessing commonsense (background) knowledge that is not expressed in the text is what we humans do in resolving references in WS sentences, and this is what we call thinking, and what probably Turing had in mind in his test for thinking machines. However, this phenomenon, of accessing background knowledge not explicitly stated in the text, is not specific to reference resolution, but is in fact the common denominator in many other linguistic phenomena.

In what follows we cover a few example sentences illustrating this phenomenon of the [missing text] - text that is never explicitly stated, but is implicitly assumed as shared background knowledge:

(14)  

a. John enjoyed the book  
⇒ John enjoyed [reading] the book  

b. John enjoyed the movie  
⇒ John enjoyed [watching] the movie

While John can, in theory, enjoy writing, publishing, buying, selling, etc. a book, and enjoy directing, producing, buying, selling, etc. a movie, a 5-year old would immediately infer that the [missing text] in (14a) is ‘reading’ and that in (14b) it is ‘watching’, and, again, precisely because these two words make the final meaning more consistent with our commonsense understanding of the world. If such examples where to be part of the WS challenge then a query posed against such sentences would be “what did John enjoy about the book” for (14a) and “what did John enjoy about the movie” for (14b).

The sentences in (15) are examples where prepositional phrase (PP) attachments must be resolved. Specifically, it must be decided what the prepositional phrase attaches to (i.e., what does it modify). Again, it seems that here humans seem to resolve these prepositional attachments by ‘discovering’ the missing text and in the process accessing the implicitly assumed and shared background knowledge.

(15) a. I read a story about evolution in ten minutes.  
⇒ I read [in 10 minutes] a story about evolution  

b. I read a story about evolution in the last million years.  
⇒ I read a story about evolution [that occurred] in the last million years.

In (15) we have are accessing background knowledge that (i) evolution is not likely to be described by 10 minutes, but the act of reading a story could. If such sentences were to be used in the WS challenge, then a good question to (15a) and (15b) would be: what is that took ten minutes/million years?

In (16) we have an example where we need to resolve what is referred to in the literature as quantifier scope ambiguities by, again, accessing the relevant background knowledge to infer the [missing text] that is not usually explicitly stated.

(16) John visited a [different] house on every street in his neighborhood.

Here the term that is often left out reflects the shared background knowledge that the physical location of a specific house is most likely to be one specific street, and it is not likely that there is a single house that is located on every street. If such questions were to be used in the WS challenge, then a good question for (16) would be: how many houses does (16) refer too (the right answers would include as one possible answer [street] - that is, the number of houses implied by (16) equals the number of streets!)

What is referred to in the literature as metonymy is yet another example of where humans use commonsense background knowledge in inferring the [missing text], as illustrated by the sentences in (17).

(17) a. The corner table wants another beer.  
⇒ The [person sitting at the] corner table wants another beer.  

b. The car in front us is annoying me, can you please pass it?  
⇒ The [person driving the] car in front us is annoying me, please pass it.

For such sentences to be part of the WS challenge, a question such as this can posed for the sentence in (17a): ‘what is the type of object that wants a beer?’ And the alternative answers would be PERSON, TABLE, and beer, for example.

There are other phenomena in natural language the understanding of which almost always requires discovering the missing text, such as copredication, nominal
compounds, etc. but the pattern is the same: like inferring the right referent for some reference, as is the case in the current WS challenge, what we usually call thinking in the process of language comprehension almost always involves discovering the missing text - text that is never explicitly stated but is assumed as shared background knowledge among a language community.

Concluding Remarks

What came to be known as the Turing Test was initially suggested by Turing as a possible answer to the question “Can Machines Think?”. As Levesque et. al. (2012) pointed out, however, the current Turing Test can be passed by programs that perform few tricks without doing anything we might call thinking. The Winograd Schema (WS) challenge, suggested by Levesque is indeed a step in the right direction, as programs correctly answering WS sentences would indeed exhibit what we might call thinking. However, in this short article we hope that we have described more formally ‘where’ appropriate (or good) WS sentences are situated. We further suggested that WS sentences are in fact instances of a much larger phenomenon. In particular, we suggested that in language compression there are many situations where, like the resolution of references in WS sentences, humans seem to employ commonsense reasoning and accessing background knowledge to infer the missing words that are not explicitly stated in the text, such as situations that involve metonymy, prepositional phrase attachments, quantifier scope ambiguity, and so on. A good test for thinking machines would therefore be a carefully selected set of several of these phenomena, all of which are examples where humans use commonsense background knowledge to infer the missing text – text not explicitly stated – but is often implicitly assumed as shared background knowledge.

References

Geman, D. et al. (2015), Visual Turing test for computer vision systems, Proceedings of the National Academy of Sciences.
Levesque, H., Davis, E. and Morgenstern, L. (2012), The Winograd Schema Challenge, In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, AAAI Press, pp. 552-561.
Shieber, S. M. (1994), Lessons from a restricted Turing Test. CACM 37(6), pp. 70–78.
Turing, A. M. (1950), Computing Machinery and Intelligence, Mind, 59(236), pp. 433-460