The Winograd Schema Challenge: Evaluating Progress in Commonsense Reasoning

Leora Morgenstern
Leidos
Autonomy and Analytics
Arlington, VA
USA
leora.morgenstern@leidos.com

Charles L. Ortiz, Jr.
Laboratory for Natural Language Processing and AI
Nuance Communications
Sunnyvale, CA
USA
charles.ortiz@nuance.com

This paper describes the Winograd Schema Challenge (WSC), which has been suggested as an alternative to the Turing Test and as a means of measuring progress in commonsense reasoning. A competition based on the WSC has been organized and announced to the AI research community. The WSC is of special interest to the AI applications community and we encourage its members to participate.

Background

Nuance Communications, Inc.is sponsoring an annual competition to encourage efforts to develop programs that can solve the Winograd Schema Challenge (WSC). The WSC was first introduced by Hector Levesque (Levesque 2011). In that paper, as well as in (Levesque, Davis, and Morgenstern 2012) and in his Research Excellence lecture at IJCAI 2013 (Levesque 2014), Levesque proposed the WSC both as an alternative to the Turing Test and as a measure for progress in commonsense reasoning. The test will be organized, administered, and evaluated by CommonsenseReasoning.org, which is dedicated to furthering research in formal commonsense reasoning .

Overview

The Turing Test is intended to serve as a test of whether a machine has achieved human-level intelligence. In one of its best-known versions (Turing 1950), a person attempts to determine whether he or she is conversing (via text) with a human or a machine. However, it has been criticized as being inadequate. At its core, the Turing Test measures a human's ability to judge deception: Can a machine fool a human into thinking that it too is human? Perhaps not surprisingly, most recent contenders for passing the Turing Test, including winners of the Loebener competition (Christian 2011) and the chatbot Eugene Goostman (University of Reading 2014), appear to be best at engaging in deceptive dialogue rather than any kind of intelligent discourse. That chatbots like Eugene Goostman can fool at least some judges into thinking they are human likely reveals more about how easy it is to fool some humans, especially in the course of a short conversation, than the bots' intelligence (Marcus 2014). Chatbots get away with evading questions that they can't answer; such evasions prevent a rigorous evaluation of a bot's ability to perform intelligent thinking.

Rather than base a test of a machine's intelligence on a short free-form conversation, Levesque's alternative envisions a test consisting of a set of multiple- choice questions that have a particular form. Three examples follow, written respectively by Levesque (2011), Ernest Davis (2012), and Terry Winograd (ostensibly in 1972). [1]

I. The trophy would not fit in the brown suitcase because it was too **big** (*small*). What was too **big** (*small*)?
Answer 0: the trophy Answer 1: the suitcase

II. My meeting started at 4:00. Since I needed to catch the train at 4:30, there wasn't much time. Luckily, it was **short** (*delayed*), so it worked out fine. What was **short** (*delayed*)?
Answer 0: the meeting Answer 1: the train

III. The town councilors refused to give the demonstrators a permit because they **feared** (*advocated*) violence. Who **feared** (*advocated*) violence?
Answer 0: the councilors Answer 1: the demonstrators

The answers to the questions (in the above examples, 0 for the sentences if the bolded words are used; 1, if the italicized words are used) are expected to be obvious to a layperson.

A human who answers these questions correctly typically uses various types of commonsense knowledge and reasoning, including his abilities in spatial, temporal, and interpersonal reasoning, and his knowledge about meetings, trains, the typical sizes of objects, and how political demonstrations unfold, to determine the correct answer. During Commonsense-2013, the Winograd Schema Challenge was therefore proposed as a promising method for tracking progress in automating commonsense reasoning.

Features of the Challenge

Winograd Schemas typically share the following features: [2]

[1] (Winograd 1972) is often cited (e.g., by (Dennett 1998; Levesque 2011; Levesque, Davis, and Morgenstern 2012)) as the source for this sentence; this is how the WSC got its name. However, we have not found the example in Winograd's book.

[2] Slightly different sets of criteria are enumerated in (Levesque 2011) and (Levesque et al., 2012). There are slight variations possible in characterizing Winograd schemas. The WSC challenge website will characterize the schema forms used in competition at least two months before the start of the competition.

1. Two (sets of) entities, not necessarily people or sentient beings, are mentioned in the sentences by noun phrases.
2. A pronoun or possessive adjective is used to reference one party (of the right sort so it can refer to either party).
3. The question involves determining the pronoun's referent.
4. There is a special word mentioned in the sentence and possibly the question. When replaced with an alternate word, the answer changes although the question still makes sense (e.g., in the examples, "big" can be changed to "small"; "feared" can be changed to "advocated".)

Significance of the WSC

Commonsense reasoning, once considered an esoteric goal left mostly to theoretical researchers (McCarthy 1986), is no longer the domain of a select group of researchers. There have been recent efforts to capture commonsense knowledge using crowdsourcing methods — e.g., recent work in Freebase and YAGO — and to encode very large commonsense knowledge bases to provide content for the semantic web. Since the 1980s, there have been efforts to develop broad coverage in formal commonsense repositories such as CYC (Lenat 1995). More recently, Virtual Personal Assistants (VPAs), which require commonsense knowledge to perform optimally, are receiving increasing attention. An ability to measure progress in commonsense reasoning is important to those engaged in the engineering of AI applications as well as those involved in basic AI research. The WSC is thus likely to be of special interest to the IAAA community.

Traditionally, research in the field of commonsense reasoning has been guided by very specific problems collectively identified by the research community as representative of targets for needed research. These have included problems in temporal reasoning, spatial reasoning, qualitative reasoning about materials, and social reasoning. However, the field has lacked the sort of challenge problems and competitions that can demonstrate the type of systematic progress found in other communities, such as machine learning or textual entailment. The WSC is the first attempt to eliminate this barrier to objectively tracking and measuring ongoing research and progress in the field.

Administration and evaluation of the test

The test, projected to consist of at least 40 Winograd Schemas, will be administered yearly, with a new set of test questions supplied each year. Ernest Davis has created more than 100 sample Winograd Schemas that can be used by participants to test their systems during development[3]. This library will be augmented yearly with the previous year's test.

Further details regarding the establishment of a baseline for human performance for each year's test, and the threshold that entries would minimally have to meet to qualify for prizes, will be available at the WSC website, **http://www.commonsensereasoning.org/winograd**. Our current plans are to grade the test in terms of the number of Winograd Schemas solved correctly. In addition, we may, at some future point, require that solutions be accompanied by a simple trace or explanation that ensures that the solution method has, in fact, demonstrated the requisite advances in commonsense reasoning. Entrants may adopt both symbolic solution approaches as well as statistical data-driven approaches. In the latter case, it will be the responsibility of the entrant to create training data that is consistent with the examples available in the library described above.

Contest rules

Individuals or teams may enter. If approved by the organizers, a team can include an industry partner. The winner that meets the baseline for human performance will receive a grand prize of $25,000. Details of other prizes will be made available at the WSC website. The current plan is to administer the test on a yearly basis starting in 2015. The first submission deadline is projected to be October 1, 2015. Additional details, including modifications to these dates, will appear at at the WSC website. A AAAI 2015 workshop, "Beyond the Turing Test," dedicated to exploring alternatives to the Turing Test, will include a discussion on methods for evaluating WSC entries. The 2015 Commonsense Reasoning Symposium, to be held at the AAAI Spring Symposium at Stanford from March 23-25, 2015, will include a special session for presentations and discussions on progress and issues related to the Winograd Schema Challenge.

Consult the WSC website for more information and updates, or send email to leora.morgenstern@leidos.com or charles.ortiz@nuance.com.

References

Christian, B. 2011. Mind vs. machine. *The Atlantic*.

Dennett, D. 1998. *Brainchildren: Essays on Designing Minds*. Bradford.

Lenat, D. B. 1995. Cyc: A large-scale investment in knowledge infrastructure. *CACM* 38(11):32–38.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Proceedings of KR 2012*.

Levesque, H. J. 2011. The Winograd Schema Challenge. In *Logical Formalizations of Commonsense Reasoning, 2011 AAAI Spring Symposium, TR SS-11-06,*.

Levesque, H. J. 2014. On our best behaviour. *Artif. Intell.* 212:27–35.

Marcus, G. 2014. What comes after the Turing Test? *New Yorker*.

McCarthy, J. 1986. Applications of circumscription to common sense reasoning. *Artif. Intell.* 28(1):89–116.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59:433–460.

University of Reading. 2014. Turing Test success marks milestone in computing history. June 8, 2014 press release.

Winograd, T. 1972. *Understanding Natural Language*. Academic Press.

[3] http://www.cs.nyu.edu/davise/papers/WS.html