Planning, Executing, and Evaluating the Winograd Schema Challenge

Leora Morgenstern, Ernest Davis, and Charles L. Ortiz, Jr.

Abstract. The Winograd Schema Challenge was proposed by Hector Levesque in 2011 as an alternative to the Turing Test. Chief among its features is a simple question format that can span many commonsense knowledge domains. Questions are chosen so that they do not require specialized knoweldge or training, and are easy for humans to answer. This paper details our plans to run the WSC and evaluate results.

Keywords: Turing Test, Winograd Schema, Pronoun Disambiguation Problem, Commonsense Knowledge, Commonsense Reasoning

The origins of the Winograd Schema Challenge

The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern, 2012) was proposed by Hector Levesque in 2011 as an alternative to the Turing Test. Turing (1950) had first introduced the notion of testing a computer system's intelligence by assessing whether it could fool a human judge into thinking that it was conversing with a human rather a computer. Although intuitively appealing and arbitrarily flexible --- in theory, a human can ask the computer system that is being tested wide-ranging questions about any subject desired --- in practice, the execution of the Turing Test turns out to be highly susceptible to systems that few people would wish to call intelligent.

The Loebner Prize Competition (Christian 2011) is in particular associated with the development of chatterbots that are best viewed as successors to ELIZA (Weizenbaum 1966), the program that fooled people into thinking that they were talking to a human psychotherapist by cleverly turning a person's statements into questions of the sort a therapist would ask. The knowledge and inference that characterize conversations of substance --- e.g., discussing alternate metaphors in sonnets of Shakespeare --- and which Turing presented as examples of the sorts of conversation that an intelligent system should be able to produce, are absent in these chatterbots. The focus is merely on engaging in surface-level conversation that can fool some humans who do not delve too deeply into a conversation, for at least a few minutes, into thinking that they are speaking to another person. The widely reported triumph of the chatterbot Eugene Goostman in fooling ten out of thirty judges to judge, after a five-minute conversation, that it was human (University of Reading 2014), was due precisely to the system's facility for this kind of shallow conversation.

Winograd Schemas

In contrast, the Winograd Schema Challenge is designed to test a system's ability to understand natural language and use commonsense knowledge. Winograd schemas (WSs) are best understood by first considering Winograd schema *halves*, which are sentences with at least one pronoun and two possible referents for that pronoun, along with a question that asks which of the two referents is correct. An example (Davis 2012) is the following:

The customer walked into the bank and stabbed one of the tellers. He was immediately taken to the emergency room.

Who was taken to the emergency room? The customer/ the teller

The correct answer is *the teller*. We know this because of all the commonsense knowledge that we have about stabbings, injuries, and how they are treated. We know that if someone is stabbed, he is very likely to be seriously wounded, and that if someone is seriously wounded, he needs medical attention. We know, furthermore, that people with acute and serious injuries are frequently treated at emergency rooms. Moreover, there is no indication in the text that the customer has been injured, and therefore no apparent reason for him to be taken to the emergency room. We reason with much of this information when we determine that the referent of "who" In the second sentence above is the teller rather than the customer.

So far, we are just describing the problem of pronoun disambiguation. Winograd schemas, however, have a twist: they are constructed so that there is a *special word* (or short phrase) that can be substituted for one of the words (or short set of words) in the sentence, causing the other candidate pronoun referent to be correct. For example, consider the above sentence with the words "police station" substituted for "emergency room":

The customer walked into the bank and stabbed one of the tellers. He was immediately taken to the police station.

Who was taken to the police station? The customer/ the teller

The correct answer now is *the police station*. To get the right answer, we use our knowledge of what frequently happens in crime scenarios --- that the alleged perpetrator is arrested and taken to the police station for questioning and booking --- together with our knowledge that stabbing someone is generally considered a crime. Since the text tells us that the customer did the stabbing, we conclude that it must be the customer, rather than the teller, who is taken to the police station.

The existence of the special word is one way to ensure that test designers do not inadvertently construct a set of problems in which ordering of words or sentence structure can be used by test takers to help the disambiguation process. For example, if a sentence with subject and object is followed by a phrase or sentence that starts with a pronoun, the subject is more likely to be the referent of the pronoun than the object. The test taker, however, who is given a Winograd schema half knows not to rely on this heuristic because the existence of the special word or set of words negates that heuristic. For instance, in the example above, *who* refers to the subject when the special set of words is *police station* but the object when the special set of words is *emergency room*.

There are three additional restrictions that we place on Winograd schemas:

First, humans should be able to easily disambiguate these questions. We are testing whether systems are as intelligent as humans, not more intelligent.

Second, they should not obey *selectional restrictions*. For example, the following would be an invalid example of a Winograd schema:

The women stopped taking the pills, because they were carcinogenic / pregnant.

What were carcinogenic / pregnant? The women/ the pills

This example is invalid because one merely needs to know that women, but not pills, can be pregnant; and that pills, but not women, can be carcinogenic; in order to solve this pronoun disambiguation problem. While this fact can also be viewed as a type of commonsense knowledge, it is generally shallower than the sort of commonsense knowledge exemplified by the emergency room / police station example above, in which one needs to reason about several commonsense facts together. The latter is the sort of deeper commonsense knowledge which we believe is characteristic of human intelligence and which we would like the Winograd Schema Challenge to test.

Third, they should be Google-proof to the extent possible. Winograd schemas should be constructed so that it is unlikely that one could use statistical properties of corpora to solve these problems.

Executing and Evaluating the Winograd Schema Challenge

When the Winograd Schema Challenge was originally conceived and developed, details of the execution of the challenge were left unspecified. In May 2013, the participants at Commonsense-2013, the Eleventh Symposium on Logical Formalizations of Commonsense Reasoning, agreed that focusing on the Winograd Schema Challenge was a high priority for researchers in commonsense reasoning. In July 2014, Nuance Communications announced its sponsorship of the Winograd Schema Challenge Competition, with cash prizes awarded for top computer systems surpassing some threshold of performance on disambiguating pronouns in Winograd schemas. At the time this article was written, the first competition was scheduled to be held at AAAI -2016 in February, 2016 in Phoenix, Arizona, assuming there are systems that are entered into competition. Because doing well at the WSC is difficult, it is possible no systems will be entered at that time; in this case, the first competition will delayed until we have received notification of interested entrants. Subsequent competitions will be held annually, biennially, or at some other set interval of time to be determined.

During the last year, we have developed a set of rules for the competition which are intended to facilitate test corpus development and participation of serious entrants. While some parts will naturally change from one competition to the next --- date and time, obviously, as well as hardware limitations --- we expect the overall structure of the competition to remain the same. Exact details are given at the Winograd Schema Challenge Competition web site at http://www.commonsensereasoning.org/winograd; the general structure and requirements are discussed below.

The competition will consist of a maximum of two rounds: a qualifying round and a final round. There will be at least sixty questions in each round. Each set of questions will have been tested on at least three human adult annotators. At least 90% of the questions in the test set will have been answered correctly by all

human annotators. The remaining questions in the test set (no more than 10% of the test set) will have been answered correctly by at least half of the human annotators. This will ensure that the questions in the test set are those for which pronoun disambiguation is easy.

It is possible that no system will progress beyond the first level, in which case the second round will not be held. The threshold required to move from the first to the second level, or to achieve a prize, must be at least 90% or no more than three percentage points below the inter-annotator agreement achieved on the test set, whichever is greater. (For example, if inter-annotator agreement on a test is 95%, the required system score is 92%.)

Prounoun Disambiguation Problems in the Winograd Schema Challenge

The first round will consist of pronoun disambiguation problems (PDPs) that are taken directly or modified from examples found in literature, biographies, autobiographies, essays, news analyses, and news stories; or have been constructed by the organizers of the competition. The second round will consist of halves of Winograd schemas; almost all of these will have been constructed by the competition organizers.

Below are some examples of the sort of pronoun disambiguation problems that could appear in the first round:

Example PDP 1:

Mrs. March gave the mother tea and gruel, while she dressed the little baby as tenderly as if it had been her own.

She dressed: Mrs. March / the mother

As if it had been: tea / gruel / baby

Example PDP 2:

Tom handed over the blueprints he had grabbed and, while his companion spread them out on his knee, walked toward the yard.

His knee: Tom/ companion

Example PDP 3:

One chilly May evening the English tutor invited Marjorie and myself into her room.

Her room: the English tutor / Marjorie

Example PDP 4:

Mariano fell with a crash and lay stunned on the ground. Castello instantly kneeled by his side and raised his head.

His head: Mariano / Castello

The following can be noted from these examples:

(1) A PDP can be taken directly from text (Example PDP 3 is taken from Vera Brittain's autobiography *Testament of Youth*) or may be modified (Examples PDP 1, 2, and 4 are modified slightly from the novels *Little Women, Tom Swift and His Airship*, and *The Pirate City: An Algerine Tale*).

(2) A pronoun disambiguation problem may consist of more than one sentence, as in Example PDP 4. In practice, we will rarely use PDPs that contain more than three sentences.

(3) There may be multiple pronouns and therefore multiple ambiguities in a sentence, as in Example PDP 1. In practice, we will have only a limited number of cases of multiple PDPs based on a single sentence or set of sentences, since misinterpreting a single text could significantly lower one's score if it is the basis for multiple PDPs.

As in Winograd schemas, a substantial amount of commonsense knowledge appears to be needed to disambiguate pronouns. For example, one way to reason that *she* in *she dressed* (Example PDP 1) refers to Mrs. March and not the mother, is to realize that the phrase "as if it were her own" implies that it (the baby) is not actually her own; that is, *she* is not the mother and must, by process of elimination, be Mrs. March. Similarly one way to understand that the English tutor is the correct referent of *her* in Example PDP 3 is through one's knowledge of the way invitations work: *X* typically invites *Y* into *X*'s domain, and not into *Z*'s domain. Especially, *X* does not invite *Y* into *Y*'s domain. Similar knowledge of etiquette come into play in Example PDP2: one way to understand that the referent of *his* is Tom is through the knowledge that *X* typically spreads documents out over *X*'s own person, and not *Y*'s person. (Other knowledge that comes into play is the fact that a person doesn't have a lap while he is walking, and the structure of the sentence entails that Tom is the individual who walks to the yard.)

Why have PDPs in the WSC Competition?

From the point of view of the computer system taking the test, there is no difference between Winograd schemas and pronoun disambiguation problems[1]. In either case, the system must choose between two (or more) possible referents for a pronoun.

Nevertheless, the move from a competition that is run solely on Winograd schemas to a competition that in its first round runs solely on pronoun disambiguation problems requires some explanation.

> [1]The first reason for having PDPs is entirely pragmatic. As originally conceived, the Winograd Schema Challenge was meant to be a one-time challenge. An example corpus of over 100 Winograd schemas was developed and published on the web (Davis 2012); Davis developed an additional 100 Winograd schemas to be used in the course of that one-time challenge. Since Nuance's decision to sponsor the Winograd Schema Challenge Competition, however, the competition is likely to be run at regular intervals, perhaps yearly. Creating Winograd schemas is difficult, requiring creativity

[1] Except that possibly there may be more than two choices in a PDP, which is disallowed in WSs by construction. So if a system notices three or more possibilities for an answer, it could know that it is dealing with a PDP. But it is a distinction without a difference; this knowledge does not seem to lead to any new approach for solution.

and inspiration, and too burdensome to do on a yearly or biennial basis.

By running the first round on PDPs, the likelihood of advancing to the second round without being able to correctly answer many of the Winograd schemas in the competition is minimized. Indeed, if a system can advance to the second round, we believe there is a good chance that it will successfully meet the Winograd Schema Challenge.

Once we had decided on using PDPs in the initial round, other advantages became apparent:

[2] Pronoun disambiguation problems occur very frequently is natural language text in the wild. One finds examples in many genres, including fiction, science fiction, biographies, and essays. In contrast Winograd schemas are fabricated natural language text and might be considered irrelevant to automated natural language processing in the real world. It is desirable to show that systems are proficient at handling the general pronoun disambiguation problem, which is a superset of the Winograd Schema Challenge. This points toward a real-world task that a system excelling in this competition should be able to do.

[3] A set of PDPs taken from the wild, and from many genres of writing, may touch on different aspects of commonsense knowledge than that which a single person or small group of people could come up with when creating Winograd schemas.

At the same time it is important to keep in mind one of the original purposes of Winograd schemas --- that the correct answer be dependent on commonsense knowledge rather than sentence structure and word order --- and to carefully choose a set of PDPs that retain this property. In addition, strong preference will be given to PDPs that do not rely on selectional restriction or on syntactical characteristics of corpora, and which are of roughly the same complexity as Winograd schemas.

Transparency

The aim of this competition is to advance science; all results obtained must be reproducible, and communicable to the public. As such, any winning entry is encouraged to furnish to the organizers of the Winograd Schema Challenge Competition its source code and executable code, and to use open source databases or knowledge bases or make its databases and knowledge structures available for independent verification of results. If an organization cannot do this, other methods for assuring reproducibility of results will be considered, such as furnishing a detailed trace of execution. Details of such methods will be published on Winograd Schema Challenge Competition web site. Entries that do not satisfy these requirements, even if excelling at the competition, will be disqualified.

An individual representing an organization's entry must be present at the competition, and must bring a laptop on which the entry will run. The specifications of the laptop to be used are given at the Winograd Schema Challenge Competition web site. It is assumed that the laptop will have a hard drive no larger than one terabyte, but researchers may negotiate this point and other details of laptop specifications with organizers. Reasonable requests will be considered.

Some entries will need to use the internet during the running of the test. This will be allowed but restricted. The room in which the competition will take place will have neither wireless nor cellular access to the internet. Internet access will be provided through a high-speed wired cable modem or fiber optic service.

Access to a highly restricted set of sites will be provided. Access to the Google search engine will be allowed. All access to the internet will be monitored and recorded.

If any entry that is eligible for a prize has accessed the internet during the competition, it will be necessary to verify that the system can achieve similar results at another undisclosed time. The laptop on which the potentially prize-winning system has run must be given to the WSCC organizers. They will then run the system on the test at some undisclosed time during a two-week period following the competition. Following the system run, organizers will compare the results obtained with the results achieved during the competition, and check that they are reasonably close. Assuming that the code contains statistical algorithms, the answers may not be identical because what is retrieved through internet query will not be exactly the same; however, the differences should be relatively small.

In the three weeks following the competition, researchers with winning or potentially winning entries will be expected to submit to WSCC organizers a paper explaining the algorithms, knowledge sources, and knowledge structures used. These papers will be posted on the commonsensereasoning.org website. Publication on the commonsensereasoning.org website does not preclude any other publication. Entries not submitting such a paper will be disqualified.

Provisional results will be announced the day after the competition. Three weeks after the competition, final results will be announced.

AI Community's Potential Gain

Publishing papers on approaches to solving the Winograd Schema Challenge is required for those eligible for a prize and highly encouraged for everyone else. All papers submitted will be posted on the Winograd Schema Challenge Competition website; it is hoped that in addition they will be submitted and published in other venues. A central aim of the Winograd Schema Challenge is that it ought to serve as motivation for research in commonsense reasoning, and we are eager to see the many directions which this research will take.

WSSC organizers will try to use the data obtained from running the competition to assess progress in automating commonsense reasoning by calculating the proportion of correct results in various subfields of commonsense reasoning. The existing example corpus and test corpus of Winograd schemas have been developed with the goal of automating commonsense reasoning, and span many areas of commonsense, including physical, spatial, and social reasoning, as well as commonsense knowledge about many common domains such as transportation, criminal acts, medical treatment, and household furnishings. PDPs will be chosen with this goal and with these areas of commonsense in mind as well.

Current plans are to annotate example PDPs and WSs with some of the commonsense areas that might prove useful in disambiguating the text. The WSCC organizers will choose an annotation scheme that is (partly) based on an existing taxonomy, such as that given by OpenCyc (http://www.opencyc.org/) or DBPedia (http://wiki.dbpedia.org/). Note that a PDP or WS might be annotated with several different commonsense domains. An entire test corpus, annotated in this way, may prove useful in assessing a system's proficiency in specific domains of commonsense reasoning. For example, a system might correctly

answer 65% of all PDPs and WSs that involve spatial reasoning; but correctly answer only 15% of all PDPs and WSs involving social reasoning. Assuming the sentences are of roughly the same complexity, this could indicate that the system is more proficient at spatial reasoning than at social reasoning.

The systems that excel in answering PDPs and WSs correctly should be capable of markedly improved natural language processing compared to current systems. For example, in translating from English to French, Google Translates often translates pronouns incorrectly, using incorrect gender, presumably because it cannot properly determine pronoun references; the technology underlying a system that wins the WSCC could improve Google Translate's performance in this regard.

More broadly, a system that contains the commonsense knowledge that facilitates correctly answering the many PDPs and WSs in competition should be capable of supporting a wide range of commonsense reasoning that would prove useful in many AI applications, including planning, diagnostics, story understanding, and narrative generation.

The sooner a system wins the Winograd Schema Challenge Competition, the sooner we will be able to leverage the commonsense reasoning that such a system would support. Even before the competition is won, however, we look forward to AI research benefiting from the commonsense knowledge and reasoning abilities that researchers build into the systems that will participate in the challenge.

Acknowledgements

This paper grew out of an invited talk by the first author at the Beyond Turing Workshop organized by Gary Marcus, Francesca Rossi, and Manuela Veloso at AAAI-2016; the ideas were further developed through conversations and email with the second and third authors after the conclusion of the workshop, and during a very productive panel session on the WSC at Commonsense-2015, held as part of the AAAI Spring Symposium Series. Thanks especially to Andrew Gordon, Jerry Hobbs, Ron Keesing, Pat Langley, Gary Marcus, and Bob Sloane for helpful discussions.

References

Christian, B. 2011 Mind vs. Machine, The Atlantic, March.

Davis, E. 2012 A Collection of Winograd Schemas, 2012, at https://www.cs.nyu.edu/davise/papers/WS.html.

Levesque, H.; Davis, E.; and Morgenstern, L. 2012 The Winograd Schema Challenge, KR-2012.

Morgenstern, L. and Ortiz, C.L. 2015 The Winograd Schema Challenge: Evaluating Progress in Commonsense Reasoning, IAAI-2015.

Turing, A. 1950 Computing Machinery and Intelligence, Mind, 1950. LIX(236), 433-460.

University of Reading 2015Turing Test success marks milestone in computing history. June 8, 2014 press release, at http://www.reading.ac.uk/news-and-events/releases/PR583836.aspx. Retrieved July 27, 2015.

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.

Leora Morgenstern is Technical Fellow and Senior Scientist at Leidos Corporation. Her research focuses on developing innovative techniques in knowledge representation and reasoning, targeted toward deep understanding of large corpora in a wide variety of domains, including legal texts, biomedical research, and social media. She heads the Executive Committee of commonsensereasoning.org, which has run the biennial Commonsense symposium series since 1991. She received a B.A. in Mathematics from the City College of New York and a Ph.D. in Computer Science from Courant Institute of Mathematical Sciences, New York Universtiy.

Ernest Davis is professor of Computer Science at New York University. His research area is automated commonsense reasoning, particularly commonsense spatial and physical reasoning. He is the author of Representing and Acquiring Geographic Knowledge (1986), Representations of Commonsense Knowledge (1990), and Linear Algebra and Probability for Computer Science Applications (2012); and co-editor of Mathematics, Substance and Surmise: Views on the Meaning and Ontology of Mathematics (2015). He received a B.Sc. in Mathematics from MIT and a Ph.D. in Computer Science from Yale.

Charles Ortiz is Senior Principal Manager of the AI and Reasoning Group at the Nuance Natural Language and AI Laboratory. His research is in collaborative multiagent systems, knowledge representation and reasoning (causation, counterfactuals, and commonsense reasoning), and robotics (cognitive and team-based robotics). His previous positions include Director of Research in Collaborative Multi-Agent Systems at the AI Center at SRI International, Adjunct Professor at Berkeley, and Postdoctoral Research Fellow at Harvard University. He received an S.B. in Physics from MIT and a Ph.D. in Computer and Information Science from the University of Pennsylvania.

.