The First Winograd Schema Challenge at IJCAI-16

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

■ *The first Winograd Schema Challenge was held in New York, New York, as part of the International Joint Conference on Artificial Intelligence. The challenge was original conceived by Hector Levesque as an alternative to the Turing test. This report details the results of this first challenge.*

The 25th International Joint Conference on Artificial Intelligence (IJCAI-16) marked the first running of the Winograd Schema Challenge, sponsored by Nuance Communications. The Winograd Schema Challenge was originally conceived by Hector Levesque (Levesque 2011; Levesque, Davis, and Morgenstern, 2012) as an alternative to the Turing test that has clear criteria for success and doesn't rely on deception. Six systems were entered, exploiting a variety of technologies. None of the systems were able to advance from the first round to the second and final round.

The Winograd Schema Challenge is concerned with finding the referents of pronouns, or solving the pronoun disambiguation problem. Doing this correctly appears to rely on having a solid base of commonsense knowledge and the ability to reason intelligently with that knowledge. This can be seen from considering an example of a Winograd schema. Like all Winograd schemas, it consists of two halves:

(1) John took the water bottle out of the backpack so that it would be lighter.

(2) John took the water bottle out of the backpack so that it would be handy.

The referent of *it* in sentence 1 is the backpack; the referent of *it* in sentence 2 is the water bottle. A human can

Contestant	Number Correct	Percentage Correct
Patrick Dhondt, Independent Researcher	27	45%
Denis Robert, Independent Researcher	19	31.666%
Nicos Issak, Open University of Cyprus	29	48.33%
Quan Liu (1), University of Science and Technology of China	28	46.9% (48.33)*
Quan Liu (2), University of Science and Technology of China	29	48.33% (58.33)*
Quan Liu (3), University of Science and Technology of China	27	45% (58.33)*

Table 1. Competition Results.

easily figure out the referent of *it* in sentence 1 because it is commonsense knowledge that when one takes an object out of a container, the object's weight remains the same, but the container weighs less. The human can figure out the referent of *it* in sentence 2 because it is commonsense knowledge that nearly all objects are handier when they are out of, rather than in, a bulky container like a backpack. This simple example draws on commonsense concepts such as weight, containment, and convenience that intelligent people typically use during their daily lives.

Sentences 1 and 2 are nearly identical except for a pair of special words or phrases; it is the choice of the special word or phrase —- in this case lighter / handy —- that changes the referent of the pronoun. All Winograd schemas have this property: this ensures that one cannot exploit properties of the structure of a particular sentence to guess at a pronoun's referent in the absence of commonsense knowledge.

The Winograd Schema Challenge Competition consists of two tests. The first test consists of pronoun disambiguation problems, most of which have been collected from naturally occurring text in fiction or nonfiction, but for which a companion schema and associated special word or phrase are not necessarily known. An example (from *Sylvester and the Magic Pebble*) is as follows:

> [3] The donkey wished a wart on its hind leg would disappear, and it did. ["It" refers to "wart," rather than "donkey" or "leg".]

The second test contains randomly chosen halves of Winograd Schemas. A system takes the second test only if it does sufficiently well on the first test. If a system can pass both tests with a mark of at least 90 percent and no less than 5 percent worse than human performance, it is eligible to win the challenge prize of $25,000. The competition has been divided into two rounds because it is more difficult to create Winograd schemas manually than to collect pronoun disambiguation problems.

There were six systems entered into the 2016 competition, representing four different teams. Table 1 summarizes their results. The asterisks for Quan Liu's three systems are due to a problem with unexpected punctuation in XML input and that affected a handful of questions. The starred scores represent performance on the corrected XML input files.

No team did well enough on the first test to qualify for the second test, so the second test was not given. The list of problems was posted on the Commonsense Reasoning website.[1] The problems on both parts of the competition were validated on human subjects in advance. The human subjects achieved better than 90 percent accuracy.

The next Winograd Schema Challenge will take place at AAAI 2018. Further information will be available on the Commonsense Reasoning website.

Notes

1. www.commonsensereasoning.org.

References

Levesque, Hector J. 2011. The Winograd Schema Challenge. In Logical Formalizations of Commonsense Reasoning: Papers from the 2011 AAAI Spring Symposium. Technical Report SS-11-06, 63–68.

Levesque, H. J.; Davis, E.; Morgenstern, L. The Winograd Schema Challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference* (KR 2012), 552–561. Palo Alto, CA: AAAI Press.

Ernest Davis is a professor at the Courant Institute of Mathematical Sciences, New York University, New York, New York.

Leora Morgenstern is a principal research scientist and technical fellow at Leidos in Arlington, Virginia.

Charles L. Ortiz is a scientist at the Laboratory for Natural Language Processing and AI at Nuance Communications in Sunnyvale, California.