DOI:10.1145/3122814

Answering questions correctly from standardized eighth-grade science tests is itself a test of machine intelligence.

BY CARISSA SCHOENICK, PETER CLARK, OYVIND TAFJORD, PETER TURNEY, AND OREN ETZIONI

Moving Beyond the Turing Test with the Allen AI Science Challenge

THE FIELD OF artificial intelligence has made great strides recently, as in AlphaGo's victories in the game of Go over world champion South Korean Lee Sedol in March 2016 and top-ranked Chinese Go player Ke Jie in May 2017, leading to great optimism for the field. But are we really moving toward smarter machines, or are these successes restricted to certain classes of problems, leaving others untouched? In 2015, the Allen Institute for Artificial Intelligence (AI2) ran its first Allen AI Science Challenge, a competition to test machines on an ostensibly difficult task—answering eighth-grade science questions. Our motivations were to encourage the field to set its sights more broadly by exploring a problem that appears to require modeling,

» key insights

- Determining whether a system truly displays artificial intelligence is difficult and complex, and well-known assessments like the Turing Test are not suited to the task.
- The Allen Institute for Artificial Intelligence suggests that answering science exam questions successfully is a better measure of machine intelligence and designed a global competition to engage the research community in this approach.
- The outcome of the Allen AI Science Challenge highlights the current limitations of AI research in language understanding, reasoning, and commonsense knowledge; the highest scores are still limited to the capabilities of information-retrieval methods.

PHOTO BY PANITAN PHOTO. WITH ROBOT ILLUSTRATION BY PETER CROWTHER ASSOCIATES

reasoning, language understanding, and commonsense knowledge in order to probe the state of the art while sowing the seeds for possible future breakthroughs.

Challenge problems have historically played an important role in motivating and driving progress in research. For a field striving to endow machines with intelligent behavior (such as language understanding and reasoning), challenge problems that test such skills are essential.

In 1950, Alan Turing proposed the now well-known Turing Test as a possible test of machine intelligence: If a system can exhibit conversational behavior that is indistinguishable from that of a human during a conversation, that system could be considered intelligent.[11] As the field of AI has grown, the test has become less meaningful as a challenge task for several reasons. First, in its details, it is not well defined (such as Who is the person giving the test?). A computer scientist would likely know good distinguishing questions to ask, while a random member of the general public may not. What constraints are there on the interaction? What guidelines are provided to the judges? Second, recent Turing Test competitions have shown that, in certain formulations, the test itself is gameable; that is, people can be fooled by systems that simply retrieve sentences and make no claim of being intelligent.[2,3] John Markoff of *The New York Times* wrote that the Turing Test is more a test of human gullibility than machine intelligence. Finally, the test as originally conceived is pass/fail rather than scored, thus providing no measure of progress toward a goal, something essential for any challenge problem.[a,b]

Machine intelligence today is viewed less as a binary pass/fail attribute and

a Turing himself did not conceive of the Turing Test as a challenge problem to drive the field forward but rather as a thought experiment to explore a useful alternative to the question Can machines think?

b Although one can imagine metrics that quantify performance on the Turing Test, the imprecision of the task definition and human variability make it difficult to define metrics that are reliably reproducible.

more as a diverse collection of capabilities associated with intelligent behavior. Rather than a single test, cognitive scientist Gary Marcus of New York University and others have proposed the notion of series of tests—a Turing Olympics of sorts—that could assess the full gamut of AI, from robotics to natural language processing.[9,12]

Our goal with the Allen AI Science Challenge was to operationalize one such test—answering science-exam questions. Clearly, the Science Challenge is not a full test of machine intelligence but does explore several capabilities strongly associated with intelligence—capabilities our machines need if they are to reliably perform the smart activities we desire of them in the future, including language understanding, reasoning, and use of commonsense knowledge. Doing well on the challenge appears to require significant advances in AI technology, making it a potentially powerful way to advance the field. Moreover, from a practical point of view, exams are accessible, measurable, understandable, and compelling.

One of the most interesting and appealing aspects of science exams is their graduated and multifaceted nature; different questions explore different types of knowledge, varying substantially in difficulty, especially for a computer. There are questions that are easily addressed with a simple fact lookup, like this

How many chromosomes does the human body cell contain?
(A) 23
(B) 32
(C) 46
(D) 64

Then there are questions requiring extensive understanding of the world, like this

City administrators can encourage energy conservation by
(A) lowering parking fees
(B) building larger parking lots
(C) decreasing the cost of gasoline
(D) lowering the cost of bus and subway fares

This question requires the knowledge that certain activities and incentives result in human behaviors that in turn result in more or less energy being consumed. Understanding the question also requires the system being able to recognize that "energy" in this context refers to resource consumption for the purposes of transportation, as opposed to other forms of energy one might find in a science exam (such as electrical and kinetic/potential).

AI vs. Eighth Grade

To put this approach to the test, AI2 designed and hosted The Allen AI Science Challenge, a four-month-long competition in partnership with Kaggle (https://www.kaggle.com/) that began in October 2015 and concluded in February 2016.[7] Researchers worldwide were invited to build AI software that could answer standard eighth-grade multiple-choice science questions. The competition aimed to assess the state of the art in AI systems utilizing natural language understanding and knowledge-based reasoning; how accurately the participants' models could answer the exam questions would serve as an indicator of how far the field has come in these areas.

Participants. A total of 780 teams participated during the model-building phase, with 170 of them eventually submitting a final model. Participants were required to make the code for their models available to AI2 at the close of the competition to validate model performance and confirm they followed contest rules. At the conclusion of the competition, the winners were also expected to make their code open source. The three teams achieving the highest scores on the challenge's test set received prizes of $50,000, $20,000, and $10,000, respectively.

Data. AI2 licensed a total of 5,083 eighth-grade multiple-choice science questions from providing partners for the purposes of the competition. All questions were standard multiple-choice format, with four answer options, as in the earlier examples. From this collection, we provided participants with a set of 2,500 training questions to train their models. We used a validation set of 8,132 questions during the course of the competition for confirming model performance. Only 800 of the validation questions were legitimate; we artificially generated the rest to disguise the real questions in order to prevent cheating via manual question answering or unfair advantage of additional training examples. A week before the end of the competition, we provided the final test set of 21,298 questions (including the validation set) to participants to use to produce a final score for their models, of which 2,583 were legitimate. We licensed the data for the competition from private assessment-content providers that did not wish to allow the use of their data beyond the constraints of the competition, though AI2 made some subsets of the questions available on its website http://allenai.org/data.

Baselines and scores. As these questions were all four-way multiple choice, a standard baseline score using random guessing was 25%. AI2 also generated a baseline score using a Lucene search over the Wikipedia corpus, producing scores of 40.2% on the training set and 40.7% on the final test set. The final results of the competition was quite close, with the top three teams achieving scores with a spread of only 1.05%. The highest score was 59.31%.

First Place

Top prize went to Chaim Linhart of Hod HaSharon, Israel (Kaggle data science website https://www.kaggle.com username Cardal). His model achieved a final score of 59.31% correct on the test question set of 2,583 questions using a combination of 15 gradient-boosting models, each with a different subset of features. Unlike the other winners' models, Linhart's model predicted the correctness of each answer option individually. Linhart used two general categories of features to make these predictions; the first consisted of information-retrieval-based features, applied by searching over corpora he compiled from various sources (such as study-guide or quiz-building websites, open source textbooks, and Wikipedia). His searches used various weightings and stemmed words to optimize performance. The other flavor of features used in his ensemble of 15 models was based on properties of the questions themselves (such as length of question and answer, form of answer like numeric answer options, answers containing referential clauses like "none of the above" as an option, and relationships among answer options).

Linhart explained that he used several smaller gradient-boosting models instead of one big model to maximize diversity. One big model tends to ignore some important features because it requires a very large training set to ensure it pays attention to all potentially useful features present. Linhart's use of several small models required that the learning algorithm use features it would otherwise ignore, an advantage, given the relatively limited training data available in the competition.

The information-retrieval-based features alone could achieve scores as high as 55% by Linhart's estimation. His question-form features filled in some remaining gaps to bring the system up to approximately 60% correct. He combined his 15 models using a simple weighted average to yield the final score for each choice. He credited careful corpus selection as one of the primary elements driving the success of his model.

In the end, each of the winning models gained from information-retrieval-based methods, indicative of the state of AI technology in this area of research.

Second Place

The second-place team, with a score of 58.34%, was from a social-media-analytics company based in Luxembourg called Talkwalker (https://www.talkwalker.com), led by Benedikt Wilbertz (Kaggle username poweredByTalkwalker).

The Talkwalker team built a relatively large corpus compared to other winning models, using 180GB of disk space after indexing with Lucene. Feature types included information-retrieval-based features, vector-based features (scoring question-answer similarity by comparing vectors from word2vec, a two-layer neural net that processes text, and GloVe, an unsupervised learning algorithm (for obtaining vector representations for words), pointwise mutual information features (measured between the question and target answer, calculated on the team's large corpus), and string hashing features in which term-definition pairs were hashed and a supervised learner was then trained to classify pairs as correct or incorrect. A final model used them to learn pairwise ranking between the answer options using the XGBoost library, an implementation of gradient-boosted decision trees.

Wilbertz's use of string hashing features was unique, not tried by either of the other two winners nor currently used in AI2's Project Aristo. His team used a corpus of terms and definitions obtained from an educational-flashcard-building site, then created negative examples by mixing terms with random definitions. A supervised classifier was trained on these incorrect pairs, and the output was used to generate features for input to XGBoost.

Third Place

The third-place winner was Alejandro Mosquera from Reading, U.K. (Kaggle username Alejandro Mosquera), with a score of 58.26%. Mosquera approached the challenge as a three-way classification problem for each pair of answer options. He transformed answer choices A, B, C, and D to all 12 possible pairs (A,B), (A,C), ..., (D,C) he labeled with three classes: left-pair element is correct; right is correct; or neither is correct. He then classified the pairs using logistic regression. This three-way classification is easier for supervised learning algorithms than the more natural two-way (correct vs. incorrect) classification with four choices, because the two-way classification requires an absolute decision about a choice, whereas the three-way classification requires only a relative ranking of the choices. Mosquera made use of three types of features: information-retrieval-based features based on scores from Elastic Search using Lucene over a corpus; vector-based features that measured question-answer similarity by comparing vectors from word2vec; and question-form features that considered such aspects of the data as the structure of a question, length of a question, and answer choices. Mosquera also noted that careful corpus selection was crucial to his model's success.

Lessons

In the end, each of the winning models gained from information-retrieval-based methods, indicative of the state of AI technology in this area of research. AI researchers intent on creating a machine with human-like intelligence are unable to ace an eighth-grade science exam because they do not currently have AI systems able to go beyond surface text to a deeper understanding of the meaning underlying each question, then use reasoning to find the appropriate answer. All three winners said it was clear that applying a deeper, semantic level of reasoning with scientific knowledge to the questions and answers would be the

key to achieving scores of 80% and higher and demonstrating what might be considered true artificial intelligence.

A few other example questions each of the top three models got wrong highlight the more interesting, complex nuances of language and chains of reasoning an AI system must be able to handle in order to answer the following questions correctly and for which information-retrieval methods are not sufficient:

What do earthquakes tell scientists about the history of the planet?

(A) Earth's climate is constantly changing.

(B) The continents of Earth are continually moving.

(C) Dinosaurs became extinct about 65 million years ago.

(D) The oceans are much deeper today than millions of years ago.

This involves the causes behind earthquakes and the larger geographic phenomena of plate tectonics and is not easily solved by looking up a single fact. Additionally, other true facts appear in the answer options ("Dinosaurs became extinct about 65 million years ago.") but must be intentionally identified and discounted as incorrect in the context of the question.

Which statement correctly describes a relationship between the distance from Earth and a characteristic of a star?

(A) As the distance from Earth to the star decreases, its size increases.

(B) As the distance from Earth to the star increases, its size decreases.

(C) As the distance from Earth to the star decreases, its apparent brightness increases.

(D) As the distance from Earth to the star increases, its apparent brightness increases.

This requires general commonsense-type knowledge of the physics of distance and perception, as well as the semantic ability to relate one statement to another within each answer option to find the right directional relationship.

Other Attempts

While numerous question-answering systems have emerged from the AI community, none has addressed the challenges of scientific and commonsense reasoning required to successfully answer these example questions. Question-answering systems developed for the message-understanding conferences[6] and text-retrieval conferences[13] have historically focused on retrieving answers from text, the former from newswire articles, the latter from various large corpora (such as the Web, microblogs, and clinical data). More recent work has focused on answer retrieval from structured data (such as "In which city was Bill Clinton born?" from FreeBase, a large publicly available collaborative knowledgebase).[4,5,15] However, these systems rely on the information being stated explicitly in the underlying data and are unable to perform the reasoning steps that would be required to conclude this information from indirect supporting evidence.

A few systems attempt some form of reasoning: Wolfram Alpha[14] answers mathematical questions, providing they are stated either as equations or with relatively simple English; Evi[10] is able to combine facts to answer simple questions (such as "Who is older: Barack or Michelle Obama?"); and START,[8] which likewise is able to answer simple inference questions (such as "What South American country has the largest population?") using Web-based databases. However, none of them attempts the level of complex question processing and reasoning that is indeed required to successfully answer many of the science questions in the Allen AI Challenge.

Looking Forward

As the 2015 Allen AI Science Challenge demonstrated, achieving a high score on a science exam requires a system that can do more than sophisticated information retrieval. Project Aristo at AI2 is focused on the problem of successfully demonstrating artificial intelligence using standardized science exams, developing an assortment of approaches to address the challenge. AI2 plans to release additional datasets and software for the wider AI research community in this effort.[1]

References

1. Allen Institute for Artificial Intelligence. Datasets; http://allenai.org/data
2. Aron, J. Software tricks people into thinking it is human. *New Scientist 2829* (Sept. 6, 2011).
3. BBC News. Computer AI passes Turing Test in 'world first.' *BBC News* (June 9, 2014); http://www.bbc.com/news/technology-27762088
4. Berant, J., Chou, A., Frostig, R., and Liang, P. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, WA, Oct. 18–21). Association for Computational Linguistics, Stroudsburg, PA, 2013, 6.
5. Fader, A., Zettlemoyer, L., and Etzioni, O. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, Aug. 24–27). ACM Press, New York, 2014.
6. Grishman, R. and Sundheim, B. Message understanding Conference-6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics* (Copenhagen, Denmark, Aug. 5–9). Association for Computational Linguistics, Stroudsburg, PA, 1996, 466–471.
7. Kaggle. *The Allen AI Science Challenge*; https://www.kaggle.com/c/the-allen-ai-science-challenge
8. Katz, B., Borchardt, G., and Felshin, S. Natural language annotations for question answering. In *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference* (Melbourne Beach, FL, May 11–13). AAAI Press, Menlo Park, CA, 2006.
9. Marcus, G., Rossi, F., and Veloso, M., Eds. Beyond the Turing Test. *AI Magazine (Special Edition) 37*, 1 (Spring 2016).
10. Simmons, J. True Knowledge: The natural language question answering Wikipedia for facts. *Semantic Focus* (Feb. 26, 2008); http://www.semanticfocus.com/blog/entry/title/true-knowledge-the-natural-language-question-answering-wikipedia-for-facts/
11. Turing, A.M. Computing machinery and intelligence. *Mind 59*, 236 (Oct. 1950), 433–460.
12. Turk, V. The plan to replace the Turing Test with a 'Turing Olympics.' *Motherboard* (Jan. 28, 2015); https://motherboard.vice.com/en_us/article/the-plan-to-replace-the-turing-test-with-a-turing-olympics
13. Voorhees, E. and Ellis, A., Eds. In *Proceedings of the 24th Text REtrieval Conference* (Gaithersburg, MD, Nov. 17–20). Publication SP 500-319, National Institute of Standards and Technology, Gaithersburg, MD, 2015.
14. Wolfram, S. Making the world's data computable. *Stephen Wolfram Blog* (Sept. 24, 2010); http://blog.stephenwolfram.com/2010/09/making-the-worlds-data-computable/
15. Yao, X. and Van Durme, B. Information extraction over structured data: Question answering with Freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Baltimore, MD, June 22–27). Association for Computational Linguistics, Stroudsburg, PA, 2014, 956–966.

Carissa Schoenick (carissas@allenai.org) is the senior program manager for Project Aristo at the Allen Institute for Artificial Intelligence in Seattle, WA.

Peter Clark (peterc@allenai.org) is the senior research manager for Project Aristo at the Allen Institute for Artificial Intelligence in Seattle, WA.

Oyvind Tafjord (oyvindt@allenai.org) is a senior research scientist and engineer at the Allen Institute for Artificial Intelligence in Seattle, WA.

Peter Turney (peter.turney@gmail.com) was a senior research scientist for Project Aristo at the Allen Institute for Artificial Intelligence in Seattle, WA, and is now retired.

Oren Etzioni (orene@allenai.org) is the Chief Executive Officer of the Allen Institute for Artificial Intelligence in Seattle, WA, and a professor in the Allen School for Computer Science at the University of Washington in Seattle, WA.

Watch the authors discuss their work in this exclusive *Communications* video. https://cacm.acm.org/videos/moving-beyond-the-turing-test