Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge!

Peter Clark
Allen Institute for AI, 2157 North Northlake Way, Seattle, WA 98103
peterc@allenai.org

Abstract

While there has been an explosion of impressive, data-driven AI applications in recent years, machines still largely lack a deeper understanding of the world to answer questions that go beyond information explicitly stated in text, and to explain and discuss those answers. To reach this next generation of AI applications, it is imperative to make faster progress in areas of knowledge, modeling, reasoning, and language. Standardized tests have often been proposed as a driver for such progress, with good reason: Many of the questions require sophisticated understanding of both language and the world, pushing the boundaries of AI, while other questions are easier, supporting incremental progress. In Project Aristo at the Allen Institute for AI, we are working on a specific version of this challenge, namely having the computer pass Elementary School Science and Math exams. Even at this level there is a rich variety of problems and question types, the most difficult requiring significant progress in AI. Here we propose this task as a challenge problem for the community, and are providing supporting datasets. Solutions to many of these problems would have a major impact on the field so we encourage you: Take the Aristo Challenge!

Introduction

While there has been an explosion of impressive, data-driven AI applications in recent years, there is still a strong need for machines that exhibit a deeper understanding of the world to support reasoning, explanation, and genuine dialog with the user. Such capabilities would open up tremendous new opportunities for real-world applications, for example in education, medicine, and scientific discovery. However, we are still far from this reality, and the recent successes of data-centric methods has at times served to distract from, rather than promote, such advances. To create this next generation of AI systems, there is a critical need for progress in the areas of knowledge, modeling, reasoning, and language.

In this paper we present a challenge task to help refocus on this longer-term goal, namely performance on Elementary School Science and Math Exams. While the task itself is not an application in its own right, attaining a high level of performance requires solving significant AI problems involving language understanding and world modeling, thus contributing to the next generation of knowledgeable AI applications. In addition, it has all the basic requirements of a good challenge problem: it is accessible, easily comprehensible, clearly measurable, and offers a graduated progression from simple tasks to those requiring deep understanding of the world. Of course, some might argue that existing techniques (large corpus statistics, deep learning, etc.) are all that is needed for even the complex questions posed in these tests. If so, we encourage you also to prove it, and take the Aristo challenge!

Fourth Grade Science and Math as a Challenge Area

Standardized tests have often been proposed a challenge problem for AI (e.g., Brachman, 2005; Fujita et al., 2014), as they appear to require significant advances in AI technology while also being accessible, measurable, understandable, and motivating. We have chosen to focus on Elementary Grade Tests (for 6-11 year olds) because the basic language processing requirements are surmountable, while the questions still present formidable challenges for solving. Similarly, we propose to focus on Science and Math to provide some initial bounds on the task. These constraints help to make the task "ambitious but realistic", although we note other groups are attempting more advanced exams, e.g., the Tokyo Entrance Exam (Strickland, 2013). We also stipulate that the exams are taken exactly as written (no reformulation or rewording),

so that the task is clear. Finally we propose to use Standardized Tests, rather than synthetic tests such as the Winograd Schema (Levesque et al., 2013) or MCTest (Richardson et al., 2013), as they provide a natural sample of problems, and more directly suggest real-world applications in the areas of education and science.

The New York Regents Science Exams: A Short Guided Tour

One of the most interesting and appealing aspects of Elementary Science exams is their graduated and multi-faceted nature: Different questions explore different types of knowledge and vary substantially in difficulty (for a computer), from a simple lookup to those requiring extensive understanding of the world. This allows incremental progress while still demanding significant advances for the most difficult questions. Information retrieval and bag-of-words methods work well for a subset of questions but eventually reach a limit, leaving a collection of questions requiring deeper understanding. We illustrate some of this variety here, using the multiple choice part of the NY Regents 4th Grade Science exams (NYSED, 2014). For a more detailed analysis, see (Clark et al., 2013).

Basic Questions

Part of the NY Regents exam tests for relatively straightforward knowledge, such as taxonomic ("isa") knowledge, definitional (terminological) knowledge, and basic facts about the world. Example questions include:

> The movement of soil by wind or water is called (A) condensation (B) evaporation (C) erosion (D) friction

> Which part of a plant produces the seeds? (A) flower (B) leaves (C) stem (D) roots

This style of question is amenable to solution by information retrieval methods and/or use of existing ontologies or fact databases, coupled with linguistic processing.

Simple Inference

Many questions are unlikely to have answers explicitly written down anywhere, from questions requiring a relatively simple leap from what might be already known to questions requiring complex modeling and understanding. An example requiring (simple) inference is:

> Which example describes an organism taking in nutrients? (A) dog burying a bone (B) A girl eating an apple (C) An insect crawling on a leaf (D) A boy planting tomatoes in the garden

Answering this question requires knowledge that eating involves taking in nutrients, and that an apple contains nutrients.

More Complex World Knowledge

Many questions appear to require both richer knowledge of the world, and appropriate linguistic knowledge to apply it to a question. As an example, consider the question below:

> Fourth graders are planning a roller-skate race. Which surface would be the best for this race? (A) gravel (B) sand (C) blacktop (D) grass

Strong correlations between sand and surface, grass and race, and gravel and graders (road smoothing machines), throw off information retrieval-based guesses. Rather, a more reliable answer requires knowing that a roller-skate race involves roller skating, that roller skating is on a surface, that skating is best on a smooth surface, and that blacktop is smooth. Obtaining these fragments of world knowledge and integrating them correctly is a substantial challenge.

As a second example, consider the question:

> A student puts two identical plants in the same type and amount of soil. She gives them the same amount of water. She puts one of these plants near a sunny window and the other in a dark room. This experiment tests how the plants respond to (A) light (B) air (C) water (D) soil

Again, information retrieval methods and word correlations do poorly. Rather, a reliable answer requires recognizing a model of experimentation (perform two tasks, differing in only one condition), knowing that being near a sunny window will expose the plant to light, and that a dark room has no light in it.

Finally, consider the question:

> A student riding a bicycle observes that it moves faster on a smooth road than on a rough road. This happens because the smooth road has (A) less gravity (B) more gravity (C) less friction (D) more friction

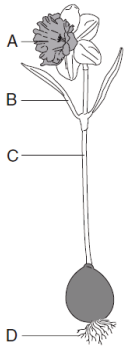
A reliable processing of this question requires envisioning and comparing two different situations, overlaying a simple qualitative model on the situations described (smoother $\rightarrow$ less friction $\rightarrow$ faster). It also requires basic knowledge that bicycles move, and that riding propels a bicycle.

Diagrams

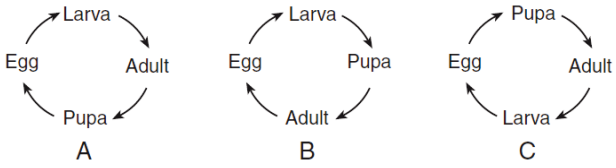
A common feature of many Elementary Grade exams is the use of diagrams in questions. We choose to include these in the challenge because of their ubiquity in tests, and because spatial interpretation and reasoning is such a

fundamental aspect of intelligence. Diagrams introduce several new dimensions to question-answering, including spatial interpretation and correlating spatial and textual knowledge. Diagrammatic (non-textual) entities in elementary exams include sketches, maps, graphs, tables, and diagrammatic representations (e.g., a food chain). Reasoning requirements include sketch interpretation, correlating textual and spatial elements, and mapping diagrammatic representations (graphs, bar charts, etc.) to a form supporting computation. Again, while there are many challenges, the level of difficulty varies widely, allowing a graduated plan of attack. Two examples are shown below, the first require sketch interpretation, part identification, and label/part correlation. The second requires recognizing and interpreting a spatial representation.

Which letter in the diagram (right) points to the plant structure that takes in water and nutrients?



Which diagram (below) correctly shows the life cycle of some insects?

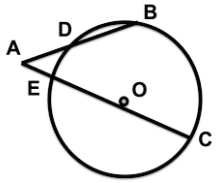


Mathematics and Geometry

As a final element, we include elementary mathematics and high school geometry questions in our challenge scope, as these questions intrinsically require mapping to mathematical models, a key requirement for many real-world tasks. These questions are particularly interesting as they combine elements of language processing, (often) story interpretation, mapping to an internal representation (e.g., algebra), and symbolic computation. For example:

Molly owns the Wafting Pie Company. This morning, her employees used 816 eggs to bake pumpkin pies. If her employees used a total of 1339 eggs today, how many eggs did they use in the afternoon?

In the diagram, AB intersects circle O at D, AC intersects circle O at E, AE = 4, AC = 24, and AB = 16. Find AD.



Such questions clearly cannot be answered by information retrieval, and instead require symbolic processing and (in the latter case) alignment of textual and diagrammatic elements (e.g., Seo et al., 2014) followed by inference.

Making the Challenge a Reality

Despite the successes of data-driven AI systems, it is imperative that we make progress in areas of knowledge, modeling, reasoning, and language if we are to make the next generation of knowledgable AI applications a reality. Elementary Grade Tests present many of these challenges, yet are also accessible, comprehensible, incremental, and easily measurable. It should be noted, though, that they do not cover all aspects of intelligence, for example spatial/kinematic reasoning, some types of commonsense reasoning, and interaction/dialog are under-represented or absent (Davis, 2014), and thus the exams do not constitute a full Turing Test; as a test of machine intelligence, they are necessary but not suffient. Nonetheless, they do cover a wide variety of problem types and levels of difficulty, making them an ideal driver for pushing the field forward. Of course, some may claim that existing data-driven techniques are all that is needed for this challenge, given enough data and computing power; if that were so, that in itself would be a startling result. Whatever your bias or philosophy, we encourage you to prove your case, and take the Aristo Challenge!

Availability: The Aristo challenge datasets are available at www.allenai.org

References

Brachman et al., "Selected Grand Challenges in Cognitive Science", MITRE Technical Report 05-1218, 2005.

Clark, P., Harrison, P., Balasubramanian, N. A Study of the Knowledge Base Requirements for Passing an Elementary Science Test. In AKBC'13, 2013.

Davis, E. The Limitations of Standardized Science Tests as Benchmarks for AI Research. NYU Technical Report, 2014. http://arxiv.org/abs/1411.1629

Fujita, A., Kameda, A., Kawazoe, A., Miyao, Y. "Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving" In Proc LREC 2014.

Levesque, H., Davis, E., Morgenstern, L. The Winograd Schema Challenge. AAAI'12, 2012.

NYSED "The Grade 4 Elementary-Level Science Test". http://www. nysedregents.org/Grade4/Science/home.html, 2014.

Richardson, M., Burges, C., Renshaw, E. MCTest: A Challenge Dataset for the Machine Comprehension of Text. EMNLP 2013.

Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O. Diagram Understanding in Geometry Questions. AAAI 2014.

Strickland, E., Can an AI Get Into the University of Tokyo? IEEE Spectrum, August 2013.