arXiv:1909.01958v3 [cs.CL] 2 Feb 2021

From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, Michael Schmitz

Allen Institute for Artificial Intelligence, Seattle, WA, U.S.A.

Abstract

AI has achieved remarkable mastery over games such as Chess, Go, and Poker, and even *Jeopardy!*, but the rich variety of standardized exams has remained a landmark challenge. Even as recently as 2016, the best AI system could achieve merely 59.3% on an 8th Grade science exam (Schoenick et al., 2016).

This article reports success on the Grade 8 New York Regents Science Exam, where for the first time a system scores more than 90% on the exam's non-diagram, multiple choice (NDMC) questions. In addition, our Aristo system, building upon the success of recent language models, exceeded 83% on the corresponding Grade 12 Science Exam NDMC questions. The results, on unseen test questions, are robust across different test years and different variations of this kind of test. They demonstrate that modern Natural Language Processing (NLP) methods can result in mastery on this task. While not a full solution to general question-answering (the questions are limited to 8th Grade multiple-choice science) it represents a significant milestone for the field.

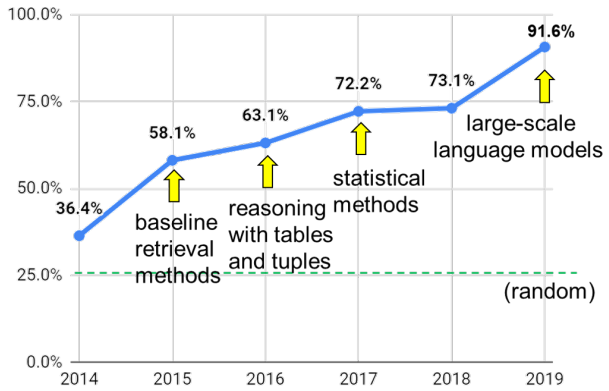

Figure 1: Progress over time of Aristo's scores on Regents 8th Grade Science (non-diagram, multiple choice questions, held-out test set).

In 2014, Project Aristo was launched with the goal of reliably answering grade-school science questions, a stepping stone in the quest for systems that understood and could reason about science. The Aristo goal was highly ambitious, with the initial system scoring well below 50% even on 4th Grade multiple choice tests. With a glance at the questions, it is easy to see why: the questions are hard. For example, consider the following 8th Grade question:

> *How are the particles in a block of iron affected when the block is melted?*
> *(A) The particles gain mass.*
> *(B) The particles contain less energy.*
> *(C) The particles move more rapidly.* **[correct]**
> *(D) The particles increase in volume.*

This question is challenging as it requires both scientific knowledge (particles move faster at higher temperatures) and common sense knowledge (melting involves raising temperature), and the ability to combine this information together appropriately.

Now, six years later, we are able to report that Aristo recently surpassed 90% on multiple choice questions from the Grade 8 New York Regents Science Exam, a major milestone and a reflection on the tremendous progress of the NLP community as a whole. In this article, we review why this is significant, how Aristo was able to achieve this score, and where the system still makes mistakes. We also explore what kinds of reasoning Aristo appears to be capable of doing, and what work still needs to be done to achieve the broader goals of the project.

Why is this an important achievement? First, passing standardized tests has been a challenging AI benchmark for many years (Bringsjord and Schimanski, 2003; Brachman et al., 2005; Strickland, 2013). A good benchmark should test a variety of capabilities while also being clearly measurable, understandable, accessible, non-gameable, and sufficiently motivating. Standardized tests, while not a full test of machine intelligence, meet many of these practical requirements (Clark and Etzioni, 2016). They also appear to require several capabilities associated with intelligence, including language understanding, reasoning, and commonsense - although the extent to which such skills are needed is controversial (Davis, 2014). We explore this in more detail this article.

Second, although NLP has made dramatic advances in recent years with the advent of large-scale language models such as ELMo (Peters et al., 2018), BERT (Devlin et al.,

2018), and RoBERTa (Liu et al., 2019), many of the demonstrated successes have been on internal yardsticks generated by the AI/NLP community itself, such as SQuAD (Rajpurkar et al., 2016), GLUE (Wang et al., 2019), and TriviaQA (Joshi et al., 2017). In contrast, the 8th Grade science exams are an external, independently-generated benchmark where we can compare machine performance with human performance. Aristo thus serves as a "poster child" for the remarkable and rapid advances achieved in NLP, applied to an easily accessible task.

Finally, Aristo makes steps towards the AI Grand Challenge of a system that can read a textbook chapter and answer the questions at the end of the chapter. This broader challenge dates back to the 1970s, and was reinvigorated in Raj Reddy's 1988 AAAI Presidential Address and subsequent writing (Reddy, 1988, 2003). However, progress on this challenge has a checkered history. Early attempts sidestepped the natural language understanding (NLU) task, in the belief that the main challenge lay in problem solving, e.g., (Larkin et al., 1980). In recent years there has been substantial progress in systems that can find factual answers in text, starting with IBM's Watson system (Ferrucci et al., 2010), and now with high-performing neural systems that can answer short questions provided they are given a text that contains the answer (e.g., Seo et al., 2016; Wang, Yan, and Wu, 2018). Aristo continues along this trajectory, but aims to also answer questions where the answer may not be written down explicitly. While not a full solution to the textbook grand challenge, Aristo is a further step along this path.

At the same time, care is needed in interpreting Aristo's results. In particular, we make no claims that Aristo is answering questions in the way a person would (and is likely using different methods). Exams are designed with human reasoning in mind, to test certain human knowledge and reasoning skills. But if the computer is answering questions in a different way, to what extent does it possess such skills? (Davis, 2014). To explore this, we examine the causes of some of Aristo's failures, and test whether Aristo has some of the semantic skills that appear necessary for good performance. We find evidence of several types of such systematic behavior, albeit not perfect, suggesting *some* form of reasoning is occurring. Although still quite distant from human problem-solving, these emergent semantic skills are likely a key contributor to Aristo's scores reaching the 90% range.

As a brief history, the metric progress of the Aristo system on the Regents 8th Grade exams (non-diagram, multiple choice part, for a hidden, held-out test set) is shown in Figure 1. The figure shows the variety of techniques attempted, and mirrors the rapidly changing trajectory of the Natural Language Processing (NLP) field in general. Early work was dominated by information retrieval, statistical, and automated rule extraction and reasoning methods (Clark et al., 2014, 2016; Khashabi et al., 2016; Khot, Sabharwal, and Clark, 2017; Khashabi et al., 2018). Later work has harnessed state-of-the-art tools for large-scale language modeling and deep learning (Trivedi et al., 2019; Tandon et al., 2018), which have come to dominate the performance of the overall system and reflects the stunning progress of the field of NLP as a whole.

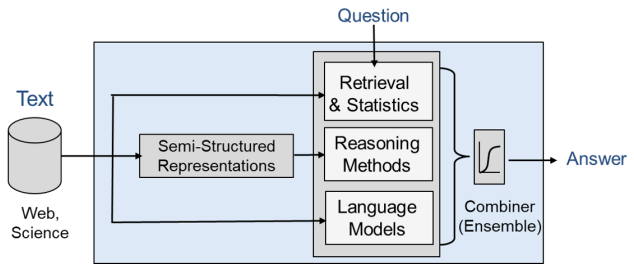

Figure 2: A simplified picture of Aristo's architecture. Aristo's eight solvers can be loosely grouped into statistical, reasoning, and language model-based approaches.

Finally, it is particularly fitting to report this result in the AI Magazine, as it is another step in the decades-long quest to fulfil the late Paul Allen's dream of a "Digital Aristotle", an "easy-to-use, all-encompassing knowledge storehouse...to advance the field of AI." (Allen, 2012), a dream also set out in the AI Magazine in the Winter 2004 edition (Friedland et al., 2004). Aristo's success reflects how much progress the field of NLP and AI as a whole has made in the intervening years.

The Aristo System

Aristo comprises of eight solvers, each of which attempts to independently answer a multiple choice question. Its suite of solvers has changed over the years, with new solvers being added and redundant solvers being dropped to maintain a simple architecture. (Earlier solvers include use of Markov Logic Networks (Khot et al., 2015), reasoning over tables (Khashabi et al., 2016), and other neural approaches, now superceded by the language models.) As illustrated in Figure 2, they can be loosely grouped into:

1. Statistical and information retrieval methods
2. Reasoning methods
3. Large-scale language model methods

We now briefly describe these solvers, with pointers to further information. Over the life of the project, the relative importance of the methods has shifted towards large-scale language methods, which now dominate the overall performance of the system.

Information Retrieval and Statistics

The **IR solver** searches to see if the question along with an answer option is explicitly stated in the corpus. To do this, for each answer option a_i, it sends q + a_i as a query to a search engine, and returns the search engine's score for the top retrieved sentence s. This is repeated for all options a_i to score them all, and the option with the highest score selected (Clark et al., 2016).

The **PMI solver** uses pointwise mutual information (Church and Hanks, 1989) to measure the associations between parts of q and parts of a_i. PMI for two n-grams x

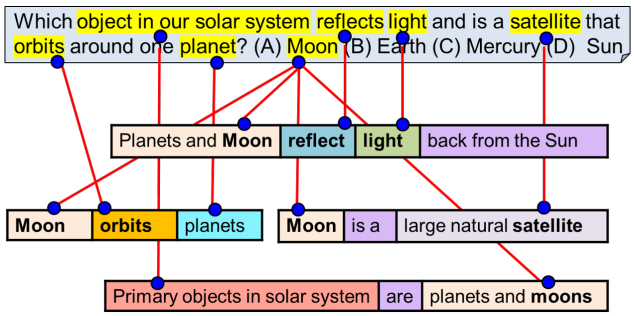

Figure 3: The TupleInference Solver retrieves tuples relevant to the question, and constructs a support graph for each answer option (Here, the support graph for the choice (A) is shown).

and y is defined as $\mathrm{PMI}(x, y) = \log \frac{p(x,y)}{p(x)p(y)}$. The larger the PMI, the stronger the association between x and y. The solver extracts unigrams, bigrams, trigrams, and skip-bigrams, and outputs the answer with the largest average PMI, calculated over all pairs of question and answer option n-grams (Clark et al., 2016).

Finally, **ACME** (Abstract-Concrete Mapping Engine) searches for a cohesive link between a question q and candidate answer a_i using a large knowledge base of *vector spaces* that relate words in language to a set of 5000 scientific terms enumerated in a *term bank*. The key insight in ACME is that we can better assess lexical cohesion of a question and answer by pivoting through scientific terminology, rather than by simple co-occurence frequencies of question and answer words (Turney, 2017).

Reasoning Methods

The **TupleInference solver** uses semi-structured knowledge in the form of *tuples*, extracted via Open Information Extraction (Open IE) (Banko et al., 2007). TupleInference treats the reasoning task as searching for a graph that best connects the terms in the question with an answer choice via the knowledge; see Figure 3 for a simple illustrative example. To find the best support graph for each answer option, we define the task as an optimization problem, and use Integer Linear Programming (ILP) to solve it. The answer choice with the highest scoring graph is then selected (Khot, Sabharwal, and Clark, 2017).

Multee (Trivedi et al., 2019) is a solver that repurposes existing *textual entailment* tools for question answering. Textual entailment (TE) is the task of assessing if one text implies another, and there are several high-performing TE systems now available. Multee learns to combine their decisions, so it can determine how strongly a set of retrieved texts entails a particular question + answer option (Trivedi et al., 2019).

The **QR (qualitative reasoning) solver** is designed to answer questions about qualitative influence, i.e., how more/less of one quantity affects another. Unlike the other solvers in Aristo, it is a specialist solver that only fires for a small subset of questions that ask about qualitative change. The solver uses a knowledge base of 50,000 (textual) statements about qualitative influence, e.g., "A sunscreen with a higher SPF protects the skin longer.". It has then been trained to reason using the BERT language model (Devlin et al., 2018), using a similar approach to that described below (Tafjord et al., 2019).

Dataset	Partition Train	Dev	Test	Total
Regents 4th	127	20	109	256
Regents 8th	125	25	119	269
Regents 12th	665	282	632	1579
ARC-Easy	2251	570	2376	5197
ARC-Challenge	1119	299	1172	2590
Totals†	4035	1151	4180	9366

†ARC (Easy+Challenge) includes Regents 4th & 8th as a subset

Table 1: Dataset partition sizes (number of questions).

Large-Scale Language Models

The field of NLP has advanced substantially with the advent of large-scale language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019). The **AristoBERT** solver applies BERT to multiple choice questions by treating the task as classification: Given a question q with answer options a_i and optional background knowledge K_i, we provide it to BERT as:

[CLS] K_i *[SEP]* q *[SEP]* a_i *[SEP]*

for each option a_i. The [CLS] output token is projected to a single logit and fed through a softmax layer across answer options, trained using cross entropy loss, and the highest scoring option selected.

AristoBERT uses three methods to apply BERT more effectively. First, we retrieve and supply background knowledge K_i along with the question when using BERT, as described above. This provides the potential for BERT to "read" that background knowledge and apply it to the question, although the exact nature of how it uses background knowledge is more complex and less interpretable. Second, following (Sun et al., 2019), we fine-tune BERT using a curriculum of several datasets, starting with RACE (a general reading comprehension dataset that is not science related (Lai et al., 2017)), followed by a collection of science training sets: OpenbookQA (Mihaylov et al., 2018), ARC (Clark et al., 2018), and Regents questions (training partition). Finally, we repeat this for three variants of BERT (cased, uncased, cased whole-word), and ensemble the predictions.

Finally, the **AristoRoBERTa** solver does the same with RoBERTa (Liu et al., 2019), a high-performing and optimized derivative of BERT trained on significantly more text. In AristoRoBERTa, we simply replace the BERT model in AristoBERT with RoBERTa, repeating similar fine-tuning steps. We ensemble two versions together, namely with and without the first fine-tuning step using RACE.

Experiments and Results

We apply Aristo to the non-diagram, multiple choice (NDMC) questions in the science exams. Although ques-

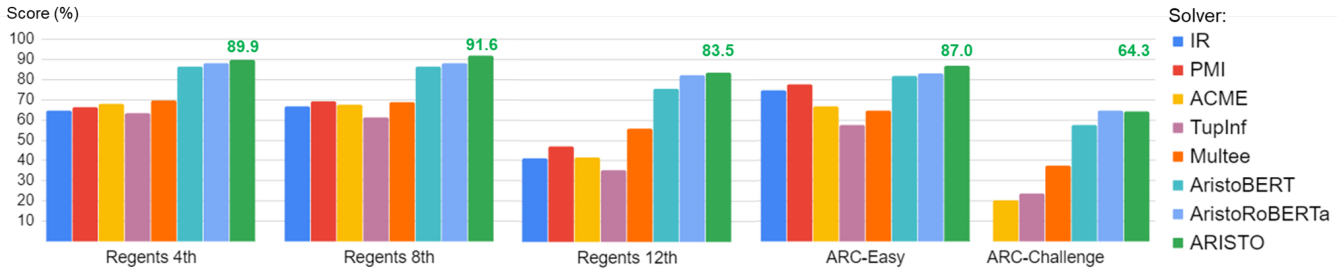

Figure 4: The results of each of the Aristo solvers, as well as the overall Aristo system, on each of the test sets. Most notably, Aristo achieves 91.6% accuracy in 8th Grade, and exceeds 83% in 12th Grade. Note that Aristo is a single system, run unchanged on each dataset (not re-tuned for each dataset).

tions with diagrams are common,[1] they are outside of our focus on language and reasoning. (For illustrative work on science diagrams, see (Krishnamurthy, Tafjord, and Kembhavi, 2016)). We also omit questions that require a direct answer for two reasons. First, after removing questions with diagrams, they are rare: For example, of the 482 direct answer questions over 13 years of Regents 8th Grade Science exams, only 38 (<8%) do *not* involve a diagram. Second, they are complex, often requiring explanation and synthesis. Both diagram and direct-answer questions are natural topics for future work.

We evaluate Aristo using the New York Regents Science exam questions[2], and also the ARC dataset, a larger corpus of science questions drawn from public resources across the country (Clark et al., 2018). The Regents exams are only produced for 4th, 8th, and 12th Grade students (corresponding to the end of elementary, middle, and high-school respectively), while the ARC questions span Grades 3 to 9. All questions are posed exactly as written, with no rewording or rephrasing. The entire dataset is partitioned into train/dev/test parts (Table 1), and for the Regents questions we ensure that each exam is either completely in train, dev, or test but not split between them. (The non-Regents ARC questions do not have exam groupings.) All but 39 of the 9366 questions are 4-way multiple choice, the remaining 39 (<0.5%) being 3- or 5-way. A random score over the entire dataset is 25.02%.

Results

The results are summarized in Figure 4, showing the performance of the solvers individually, and their combination in the full Aristo system. Note that Aristo is a single system run on the five datasets (not retuned for each dataset in turn).

Most notably, Aristo's scores on the Regents Exams far exceed earlier performances (e.g., Schoenick et al., 2016; Clark et al., 2016), and represent a new high-point on science questions.

In addition, the results show the dramatic impact of new language modeling technology, embodied in AristoBERT and AristoRoBERTa, the scores for these two solvers dominating the performance of the overall system. Even on the ARC-Challenge questions, containing a wide variety of difficult questions, the language modeling based solvers dominate. The general increasing trend of solver scores from left to right for each test set loosely reflects the progression of the NLP field over the six years of the project.

To further check that we have not overfit to our data, we also ran Aristo on the most recent years of the Regents Exams (4th and 8th Grade), years 2017-19, that were unavailable at the start of the project and were not part of our datasets. We find similar scores (average 92.8% for the three Fourth Grade exams, 93.3% for the Eighth Grade), suggesting the system is not overfit.

On a combination of exam scores and lab work (weighted approximately 60:40), the NY State Education Department considers an overall score of 65% as "Meeting the Standards", and over 85% as "Meeting the Standards with Distinction"[3]. As a somewhat loose comparison, if this rubric applies equally to the NDMC subset we have studied, this would mean Aristo has met the standard with distinction in 8th Grade NDMC Science (although clearly full science requires substantially more).

Answer Only Performance

Several authors have observed that for some multiple choice datasets, systems can still perform well even when ignoring the question body and looking only at the answer options (Gururangan et al., 2018; Poliak et al., 2018b). This surprising result is particularly true for crowdsourced datasets, where workers may use stock words or phrases (e.g., "not") in incorrect answer options that gives them away. To measure this phenomenon on our datasets, we trained and tested a new AristoRoBERTa model giving it only the answer options, i.e., no question body nor retrieved knowledge. (Without retraining the model scores slightly less, 35% overall vs. 38% with retraining). The results (test set) are shown in Figure 5, indicating that it is hard to select the right answer without reading the question. (Scores are slightly higher for

[1] Ratios of non-diagram multiple choice (NDMC), with-diagram multiple-choice (DMC), non-diagram direct answer (NDDA), and with-diagram direct answer (DDA) questions are approximately 45/25/5/25 for Regents 4th Grade, 25/25/5/45 for 8th Grade, and 40/25/15/20 for 12th grade.

[2] See https://www.nysedregents.org/ for the original exams.

[3] https://www.nysedregents.org/grade8/science/618/home.html

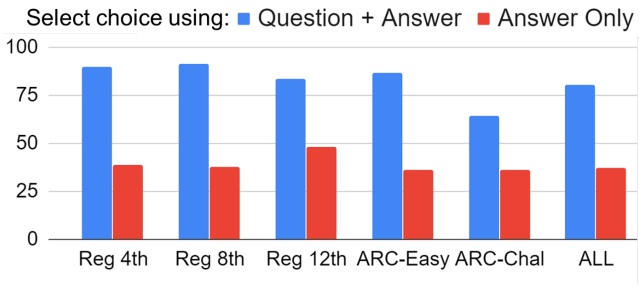

Figure 5: Scores when looking at the answer options only, compared with using the full questions. The (desirably) low scores/large drops indicate it is hard to guess the answer without reading the question.

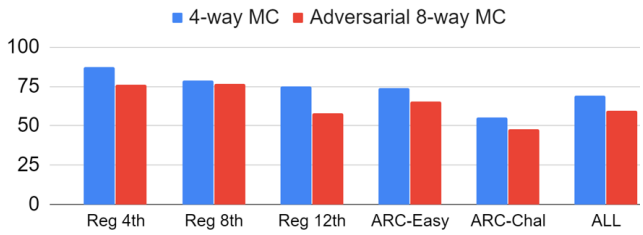

Figure 6: Aristo's scores drop a small amount (average 10%) when tested on adversarially generated 8-way MC questions.

12th Grade answer-only, possibly because the average answer length is longer, hence more potential for hidden patterns inside that hint at correctness/incorrectness).

Adversarial Answer Options

What if we add extra incorrect answer options to the questions? If a system has mastery of the material, we would expect its score to be relatively unaffected by such modifications. We can make this more challenging by doing this *adversarially*: try many different incorrect options until the system is fooled. If we do this, turning a 4-way MC question into 8-way with options chosen to fool Aristo, then retrain on this new dataset, we do observe an effect: the scores drop, although the drop is small (≈10%), see Figure 6. This indicates that while Aristo performs well, it still has some blind spots that can be artificially uncovered through adversarial methods such as this.

Analysis

Despite the high scores, Aristo still makes occasional mistakes. Because Aristo retrieves and then "reads" corpus sentences to answer a question, we can inspect the retrieved knowledge when Aristo fails, and gain some insight as to where and why it makes errors. Did Aristo retrieve the "right" knowledge, but then choose the wrong answer? Or was the failure due (in part) to the retrieval step itself? We manually analyzed 30 random failures (of 248) in the entire dev set (Regents + ARC, 1151 dev set questions total), and found four main categories of failures, illustrated in Figure 7, that we now summarize. As the language model solvers have highest weight in Aristo, we conduct this analysis for failures by AristoRoBERTa, but note these very frequently (∼90% of the time) equate to overall Aristo failures, and that when AristoRoBERTa fails, most (on average, 76%) of the other solvers also fail also. We did not discern any systematic patterns within these.

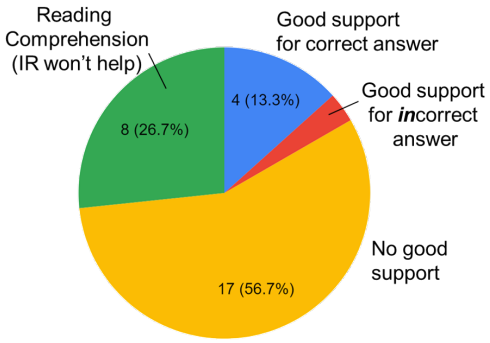

Figure 7: Case Study: Causes of error for 30 questions that Aristo answered incorrectly.

Good Support for Correct Answer (13%) Surprisingly, only 4 of the 30 failures were cases where the retrieved knowledge supported the right answer option, but Aristo chose a wrong answer option. An example is:

> *Which is the best unit to measure distances between Earth and other solar systems in the universe? (A) miles (B) kilometers (C) light years* **[correct]** *(D) astronomical units* **[selected]**

Here, although Aristo *did* retrieve good evidence for the correct answer (C), namely:

> *Distances between Earth and the stars are often measured in terms of light-years.*

it still preferred the incorrect option (D) from the retrieved knowledge:

> *In general, distances in the solar system are measured in astronomical units.*

Here, Aristo has confused distinguishing distances *within* the Solar System vs. distances *between* solar systems. (A confusion that a human might easily make too). This illustrates where Aristo has mis-applied its retrieved knowledge. However, such cases appear to be rare (4/30). In other words, for the fast majority of questions, *if* suitable knowledge is retrieved, then Aristo will answer correctly.

No Support for the Correct Answer (57%) The largest cause of failure was simply when none of the retrieved sentences provide evidence for the correct answer. In such situations, Aristo has little chance of answering correctly. For example:

Although they belong to the same family, an eagle and a pelican are different. What is one difference between them? (A) their preference for eating fish (B) their ability to fly (C) their method of reproduction **[selected]** *(D) their method of catching food* **[correct]**

As there are no corpus sentences comparing eagles and pelicans, Aristo retrieves a rather random collection of unhelpful facts. Instead, what is needed here is to realize this is a comparison question, then retrieve appropriate facts for pelicans and eagles individually, and them compare them, for example by using question decomposition methods (Wolfson et al., 2020).

Reading Comprehension (27%) In the exams, there are a few "reading comprehension" questions that primarily require reasoning over the question content itself, rather than retrieving and applying science knowledge. In such situations, retrieved knowledge is unlikely to be helpful. 8/30 failures fell into this category. One example is a question describing an experiment:

A student wants to determine the effect of garlic on the growth of a fungus species. Several samples of fungus cultures are grown in the same amount of agar and light. Each sample is given a different amount of garlic. What is the independent variable in this investigation? (A) amount of agar (B) amount of light (C) amount of garlic **[correct]** *(D) amount of growth* **[selected]**

Here, the answer is unlikely to be written down in a corpus, as a novel scenario is being described. Rather, it requires understanding the scenario itself.

A second example is a meta-question about sentiment:

Which statement is an opinion? (A) Many plants are green. (B) Many plants are beautiful. **[correct]** *(C) Plants require sunlight.* **[selected]** *(D) Plants can grow in different places.*

Again, retrieval is unlikely to help here. Rather, the question asks for an analysis of the options themselves, something Aristo does not realize.

Good Support for an Incorrect Answer (3%) Occasionally a failure occurs due to retrieved knowledge supporting an incorrect answer, e.g., if the question is ambiguous, or the retrieved knowledge is wrong. The single failure in this category that we observed was:

Which of these objects will most likely float in water? (A) glass marble (B) steel ball (C) hard rubber ball **[selected]** *(D) table tennis ball* **[correct]**

Here, Aristo retrieved evidence for both (C) and (D), e.g., for (C), Aristo's retrieval included:

> *"It had like a rubber ball in it, which would maybe float up. . . "*

here leading Aristo to select the wrong answer. Arguably, as this question is a comparative ("...which *most likely* floats?"), Aristo should have rejected this in favor of the correct answer (*table tennis ball*). However, as Aristo computes a confidence for each option independently, it is unable to directly make these cross-option comparisons.

Other Finally we point to one other interesting failure:

About how long does it take for the Moon to complete one revolution around Earth? (A) 7 days (B) 30 days **[correct]** *(C) 90 days (D) 365 days* **[selected]**

In this case, many relevant sentences were retrieved, including:

- *Because it takes the moon about* **27.3 days** *to complete one orbit around the Earth...*
- *It takes* **27.3 days** *for the moon to complete one revolution around the earth.*
- *The Moon completes one revolution around the Earth in* **27.32166** *days.*

However, Aristo does not realize 27.3 is "about" 30, and hence answered the question incorrectly.

A Score Card for Aristo's Semantic Skills

From informal tests, Aristo appears to be doing more than simply matching a question + answer option to a retrieved sentence. Rather, Aristo appears to recognize various linguistic and semantic phenomena, and respond appropriately. For example, if we add negation (a "not") to the question, Aristo almost always correctly changes its answer choice. Similarly if we replace "increase" with "decrease" in a question, Aristo will typically change its answer choice correctly, suggesting it has some latent knowledge of qualitative direction.

To quantify such skills more systematically, we performed five sets of tests on Aristo *without fine-tuning Aristo on those tests*, i.e., the tests are zero-shot. (From other experiments, we know that if we train Aristo on these tests it can perform them almost perfectly, but our interest here is how Aristo performs "out of the box" after training on the science exams). Each test probes a different semantic phenomenon of interest, as we now describe.

Negation How well does Aristo handle negation? As a (limited) test, we generated a synthetic negation dataset (10k questions), where each question has a synthetic context (replacing the retrieved sentences), plus a question about it, for example:

Context:
Alan is small. Alan is tall. Bob is big. Bob is tall. Charlie is big. Charlie is tall. David is small. David is short.
Question:
Which of the following is **not** *tall?*
(A) Alan (B) Bob (C) Charlie (D) David **[correct]**

We then test Aristo on this dataset *without fine-tuning on it*. Remarkably, Aristo score 94% on this dataset, suggesting at least in this particular formulation, Aristo has an understanding of "not".

Conjunction We test conjunction in a similar way, with questions such as:

> **Context:**
> *Alan is red. Alan is big. Bob is blue. Bob is small. Charlie is blue. Charlie is big. David is red. David is small.*
> **Question:**
> *Which of the following is big* **and** *blue? (A) Alan (B) Bob (C) Charlie* **[correct]** *(D) David*

With questions containing two conjuncts (e.g., the one above), and again *without* any fine-tuning on this data, Aristo scores 98%. If we increase the number of conjuncts in the question to 3, 4, and 5, Aristo scores 95%, 94%, and 80% respectively. If we use five conjuncts *and* a negation, for example:

> **Context:**
> *Alan is red. Alan is big. Alan is light. Alan is old. Alan is tall. Bob is red. Bob is small. Bob is heavy. Bob is old. Bob is tall. Charlie is blue. Charlie is big. Charlie is light. Charlie is old. Charlie is tall. David is red. David is small. David is heavy. David is young. David is tall.*
> **Question:**
> *Which of the following is old* **and** *red* **and** *light* **and** *big* **and not** *short? (A) Alan (B) Bob (C) Charlie (D) David*

(the correct answer is left as an exercise for the reader) Aristo remarkably still scores 75%. Although not perfect, this indicates some form of systematic handling of conjunction plus negation is occurring.

Polarity Polarity refers to Aristo's ability to correctly change its answer when a comparative in the question is "flipped". For example, given:

> *Which human activity will likely have a negative effect on global stability?*
> *(A) decreasing water pollution levels*
> *(B) increasing world population growth.* **[correct]**

if we now switch "negative" to "positive", Aristo should switch its answer from (B) to (A). To score a point, Aristo must get *both* the original question *and* the "flipped" question (with a changed answer) correct. To measure this, we use an existing qualitative dataset containing such pairs, called QuaRTz (Tafjord et al., 2019). As the QuaRTz questions are 2-way multiple-choice, a random score for getting *both* right would be 25%. Remarkably, we find Aristo scores 67.1% (again with no fine-tuning on QuaRTz), suggesting Aristo has some knowledge about the polarity of comparatives. Note that this test also requires Aristo to get the original question right in the first place, thus the score reflects both knowledge and polarity reasoning, a harder task than polarity alone.

Factuality Event factuality refers to whether an event, mentioned in a textual context, did in fact occur (Rudinger, White, and Durme, 2018). For example:

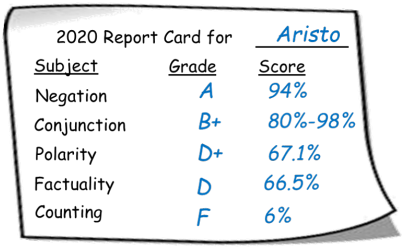

Figure 8: Aristo, with no fine-tuning, passes probes for all but counting.

> *If someone regretted that a particular thing happened*
> *(A) that thing might or might not have happened*
> *(B) that thing didn't happen*
> *(C) that thing happened* **[correct]**

Predicting factuality requires understanding what the context around an event implies about that event's occurrence. We tested Aristo on this task using the veradicity question dev set from DNC (Poliak et al., 2018a), converted to multiple choice format (394 questions). On this task, Aristo score 66.5%, again suggesting Aristo has some knowledge of how words affect the factiveness of the events that they modify.

Counting Finally we ran Aristo on bAbI task 7, (a simple form of) counting (Weston et al., 2015), converted to multiple choice with four options, as below. For example:

> *Daniel picked up the football. Daniel dropped the football. Daniel got the milk. How many objects is Daniel holding? (A) zero (B) one (C) two (D) three*

Aristo (again not fine-tuned on this dataset) did badly at this task, scoring only 6%.[4] This results is perhaps not surprising, as this type of reasoning is not exemplified in any way in an any of Aristo's training data.

Scorecard We can informally map these scores to a grade level to give Aristo a Score Card (Figure 8). The most striking observation is that Aristo "passes" all but counting, and has apparently acquired these skills through its general fine-tuning on RACE and Science Exams, with no fine-tuning at all on these specific probing datasets. Aristo does appear to be doing more than just sentence matching, but not quite in the systematic way a person would. These acquired latent skills are reflected in the high scores Aristo achieves on the Science Exams.

Discussion

What can we conclude from this? Most significantly, Aristo has achieved surprising success on a formidable problem,

[4] Lower than random guessing, because Aristo frequently selects option D ("three"), an option which is (by chance) very rarely the right answer in this dataset. "D" is likely chosen due to a small random bias towards "three", and all questions looking stylistically similar. Note there is no training on this (nor other) probing datasets, so Aristo is unaware of the answer distribution.

in particular by leveraging large-scale language models. The system thus serves as a demonstration of the stunning progress that NLP technology as a whole has made in the last two years.

At the same time, exams themselves are an imperfect test of understanding science, and, despite their many useful properties, are also only a partial test of machine intelligence (Davis, 2014). Earlier, we highlighted several classes of problems Aristo does not handle well, even though its exam scores are high: questions requiring diverse pieces of evidence to be combined, reading comprehension ("story") questions, meta-questions, and arithmetic. Davis (2016) has similarly pointed out that as standardized tests are authored for people, not machines, they also *don't* directly test for things that are easy for people such as temporal reasoning, simple counting, and obviously impossible situations. These are problem types that Aristo is not familiar with and would be hard for it (but not for people). Clearly, science exams are just one of many different, partial indicators of progress in broader AI. Finally, we have only been using multiple-choice questions, a format that just requires ranking of answer choices, arguably allowing more use of weak evidence compared with (say) generating an answer, or even independently deciding if an answer is true or false (Clark et al., 2019).

On the other hand, we *do* see clear evidence of systematic semantic skill in Aristo. For example, Aristo not only answers this question correctly:

> *City administrators can encourage energy conservation by*
> *(A) lowering parking fees*
> *(B) decreasing the cost of gasoline.*
> *(C) lowering the cost of bus and subway fares.* **[correct,selected]**

but flipping "decreasing" and "lowering" causes it to correctly change its answer:

> *City administrators can encourage energy conservation by*
> *(A) lowering parking fees*
> *(B) ~~decreasing~~* **increasing** *the cost of gasoline.* **[correct,selected]**
> *(C) ~~lowering~~* **raising** *the cost of bus and subway fares.*

Our probes showed that such behavior is not just anecdotal but systematic, suggesting that *some* form of reasoning is occurring, but not in the traditional style of discrete symbol manipulation in a formally designed language (Brachman and Levesque, 1985; Genesereth and Nilsson, 2012). Other work has similarly found that neural systems can learn systematic behavior (Lake and Baroni, 2017; Clark, Tafjord, and Richardson, 2020), and these emergent semantic skills are a key contributor to Aristo's scores reaching the 90% range. Large-scale language model architectures have brought a dramatic, new capability to the table that goes significantly beyond just pattern matching and similarity assessment.

Related Work on Standardized Testing for AI

Standardized Tests

Standardized tests have long been proposed as challenge problems for AI (e.g., Bringsjord and Schimanski, 2003; Brachman et al., 2005; Clark and Etzioni, 2016; Piatetsky-Shapiro et al., 2006), as they appear to require significant advances in AI technology while also being accessible, measurable, understandable, and motivating.

Earlier work on standardized tests focused on specialized tasks, for example, SAT word analogies (Turney, 2006), GRE word antonyms (Mohammad et al., 2013), and TOEFL synonyms (Landauer and Dumais, 1997). More recently, there have been attempts at building systems to pass university entrance exams. Under NII's Todai project, several systems were developed for parts of the University of Tokyo Entrance Exam, including maths, physics, English, and history (Strickland, 2013; Tainaka, 2013; Fujita et al., 2014), although in some cases questions were modified or annotated before being given to the systems (e.g., Matsuzaki et al., 2014). Similarly, a smaller project worked on passing the Gaokao (China's college entrance exam) (e.g., Cheng et al., 2016; Guo et al., 2017). The Todai project was reported as ended in 2016, in part because of the challenges of building a machine that could "grasp meaning in a broad spectrum" (Mott, 2016).

Math Word Problems

Substantial progress has been achieved on math word problems. On plane geometry questions, (Seo et al., 2015) demonstrated an approach that achieve a 61% accuracy on SAT practice questions. The Euclid system (Hopkins et al., 2017) achieved a 43% recall and 91% precision on SAT "closed-vocabulary" algebra questions, a limited subset of questions that nonetheless constitutes approximately 45% of a typical math SAT exam. Closed-vocabulary questions are those that do not reference real-world situations (e.g., "what is the largest prime smaller than 100?" or "Twice the product of x and y is 8. What is the square of x times y?")

Work on open-world math questions has continued, but results on standardized tests have not been reported and thus it is difficult to benchmark the progress relative to human performance. See Amini et al. (2019) for a recent snapshot of the state of the art, and references to the literature on this problem.

Summary and Conclusion

Answering science questions is a long-standing AI grand challenge (Reddy, 1988; Friedland et al., 2004). We have described Aristo, the first system to achieve a score of over 90% on the non-diagram, multiple choice part of the New York Regents 8th Grade Science Exam, demonstrating that modern NLP methods can result in mastery of this task. Although Aristo only answers multiple choice questions without diagrams, and operates only in the domain of science, it nevertheless represents an important milestone towards systems that can read and understand. The momentum on

this task has been remarkable, with accuracy moving from roughly 60% to over 90% in just three years. In addition, the use of independently authored questions from a standardized test allows us to benchmark AI performance relative to human students.

Beyond the use of a broad vocabulary and scientific concepts, many of the benchmark questions intuitively appear to require some degree of reasoning to answer. For many years in AI, reasoning was thought of as discrete symbol manipulation. With the advent of deep learning, this notion of reasoning has expanded, with systems performing challenging tasks using neural architectures rather than explicit representation languages. Similarly, we observe surprising performance on answering science questions, and on specific semantic phenomena such as those probed earlier. This suggests that the machine has indeed learned something about language and the world, and how to manipulate that knowledge, albeit neither symbolically nor discretely.

Although an important milestone, this work is only a step on the long road toward a machine that has a deep understanding of science and achieves Paul Allen's original dream of a Digital Aristotle. A machine that has fully understood a textbook should not only be able to answer the multiple choice questions at the end of the chapter - it should also be able to generate both short and long answers to direct questions; it should be able to perform constructive tasks, e.g., designing an experiment for a particular hypothesis; it should be able to explain its answers in natural language and discuss them with a user; and it should be able to learn directly from an expert who can identify and correct the machine's misunderstandings. These are all ambitious tasks still largely beyond the current technology, but with the rapid progress happening in NLP and AI, solutions may arrive sooner than we expect.

Acknowledgements

We gratefully acknowledge the late Paul Allen's inspiration, passion, and support for research on this grand challenge. We also thank the many other contributors to Aristo, including Niranjan Balasubramanian, Matt Gardner, Peter Jansen, Jayant Krishnamurthy, Souvik Kundu, Todor Mihaylov, Harsh Trivedi, Peter Turney, and the Beaker team at AI2, and to Ernie Davis, Gary Marcus, Raj Reddy, and many others for helpful feedback on this work. We also thank the anonymous reviewers for helpful comments that improved the paper.

References

Allen, P. 2012. Idea Man: A memoir by the Cofounder of Microsoft. Penguin.

Amini, A.; Gabriel, S.; Lin, P.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In NAACL-HLT.

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open Information Extraction from the Web. In IJCAI.

Brachman, R. J., and Levesque, H. J. 1985. Readings in Knowledge Representation. Morgan Kaufmann Publishers Inc.

Brachman, R.; Gunning, D.; Bringsjord, S.; Genesereth, M.; Hirschman, L.; and Ferro, L. 2005. Selected Grand Challenges in Cognitive Science. Technical report, MITRE Technical Report 05-1218. Bedford MA: The MITRE Corporation.

Bringsjord, S., and Schimanski, B. 2003. What is Artificial Intelligence? Psychometric AI as an Answer. In IJCAI, 887–893. Citeseer.

Cheng, G.; Zhu, W.; Wang, Z.; Chen, J.; and Qu, Y. 2016. Taking up the Gaokao Challenge: An Information Retrieval Approach. In IJCAI, 2479–2485.

Church, K. W., and Hanks, P. 1989. Word Association Norms, Mutual Information and Lexicography. In 27th ACL, 76–83.

Clark, P., and Etzioni, O. 2016. My Computer is an Honor Student - But how Intelligent is it? Standardized Tests as a Measure of AI. AI Magazine 37(1):5–12.

Clark, P.; Balasubramanian, N.; Bhakthavatsalam, S.; Humphreys, K.; Kinkead, J.; Sabharwal, A.; and Tafjord, O. 2014. Automatic Construction of Inference-Supporting Knowledge Bases. In 4th Workshop on Automated Knowledge Base Construction (AKBC).

Clark, P.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P. D.; and Khashabi, D. 2016. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions. In AAAI, 2580–2586.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. ArXiv abs/1803.05457.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In NAACL-HLT.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In IJCAI.

Davis, E. 2014. The Limitations of Standardized Science Tests as Benchmarks for Artificial Intelligence Research. ArXiv abs/1411.1629.

Davis, E. 2016. How to Write Science Questions that are Easy for People and Hard for Computers. AI Magazine 37:13–22.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J.; et al. 2010. Building Watson: An Overview of the DeepQA Project. AI magazine 31(3):59–79.

Friedland, N. S.; Allen, P. G.; Matthews, G.; Witbrock, M.; Baxter, D.; Curtis, J.; Shepard, B.; Miraglia, P.; Angele,

J.; Staab, S.; et al. 2004. Project Halo: Towards a Digital Aristotle. AI magazine 25(4):29–29.

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

Genesereth, M. R., and Nilsson, N. J. 2012. Logical Foundations of Artificial Intelligence. Morgan Kaufmann.

Guo, S.; Zeng, X.; He, S.; Liu, K.; and Zhao, J. 2017. Which is the Effective Way for Gaokao: Information Retrieval or Neural Networks? In EACL'17, 111–120.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In NAACL.

Hopkins, M.; Petrescu-Prahova, C.; Levin, R.; Bras, R. L.; Herrasti, A.; and Joshi, V. 2017. Beyond Sentential Semantic Parsing: Tackling the Math SAT with a Cascade of Tree Transducers. In EMNLP.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In ACL'17. Vancouver, Canada: Association for Computational Linguistics.

Khashabi, D.; Khot, T.; Sabharwal, A.; Clark, P.; Etzioni, O.; and Roth, D. 2016. Question Answering via Integer Programming over Semi-Structured Knowledge. In IJCAI.

Khashabi, D.; Khot, T.; Sabharwal, A.; and Roth, D. 2018. Question Answering as Global Reasoning over Semantic Abstractions. In AAAI.

Khot, T.; Balasubramanian, N.; Gribkoff, E.; Sabharwal, A.; Clark, P.; and Etzioni, O. 2015. Exploring Markov Logic Networks for Question Answering. In EMNLP.

Khot, T.; Sabharwal, A.; and Clark, P. F. 2017. Answering Complex Questions using Open Information Extraction. In ACL.

Krishnamurthy, J.; Tafjord, O.; and Kembhavi, A. 2016. Semantic Parsing to Probabilistic Programs for Situated Question Answering. In EMNLP.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale Reading Comprehension Dataset from Examinations. In EMNLP.

Lake, B. M., and Baroni, M. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In ICML.

Landauer, T. K., and Dumais, S. T. 1997. A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. Psychological review 104(2):211.

Larkin, J. H.; McDermott, J.; Simon, D. P.; and Simon, H. A. 1980. Models of Competence in Solving Physics Problems. Cognitive Science 4:317–345.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Matsuzaki, T.; Iwane, H.; Anai, H.; and Arai, N. H. 2014. The most Uncreative Examinee: a First Step toward Wide Coverage Natural Language Math Problem Solving. In AAAI'14.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In EMNLP.

Mohammad, S. M.; Dorr, B. J.; Hirst, G.; and Turney, P. D. 2013. Computing Lexical Contrast. Computational Linguistics 39(3):555–590.

Mott, N. 2016. Todai Robot Gives Up on Getting Into the University of Tokyo. Inverse. (https://www.inverse.com/article/23761-todai-robot-gives-up-university-tokyo).

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M. P.; Clark, C.; Lee, K.; and Zettlemoyer, L. S. 2018. Deep Contextualized Word Representations. In NAACL.

Piatetsky-Shapiro, G.; Djeraba, C.; Getoor, L.; Grossman, R.; Feldman, R.; and Zaki, M. 2006. What are the Grand Challenges for Data Mining?: KDD-2006 Panel Report. ACM SIGKDD Explorations Newsletter 8(2):70–77.

Poliak, A.; Haldar, A.; Rudinger, R.; Hu, J. E.; Pavlick, E.; White, A. S.; and Durme, B. V. 2018a. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In EMNLP.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018b. Hypothesis Only Baselines in Natural Language Inference. In StarSem.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP.

Reddy, R. 1988. Foundations and Grand Challenges of Artificial Intelligence: AAAI Presidential Address. AI Magazine 9(4).

Reddy, R. 2003. Three Open Problems in AI. J. ACM 50:83–86.

Rudinger, R.; White, A. S.; and Durme, B. V. 2018. Neural Models of Factuality. In NAACL-HLT.

Schoenick, C.; Clark, P. F.; Tafjord, O.; Turney, P. D.; and Etzioni, O. 2016. Moving beyond the Turing Test with the Allen AI Science Challenge. CACM.

Seo, M. J.; Hajishirzi, H.; Farhadi, A.; Etzioni, O.; and Malcolm, C. 2015. Solving Geometry Problems: Combining Text and Diagram Interpretation. In EMNLP.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional Attention Flow for Machine Comprehension. ArXiv abs/1611.01603.

Strickland, E. 2013. Can an AI get into the University of Tokyo? IEEE Spectrum 50(9):13–14.

Sun, K.; Yu, D.; Yu, D.; and Cardie, C. 2019. Improving Machine Reading Comprehension with General Reading Strategies. In NAACL-HLT.

Tafjord, O.; Gardner, M.; Lin, K.; and Clark, P. 2019. QuaRTz: An Open-Domain Dataset of Qualitative Relationship Questions. In EMNLP.

Tainaka, M. 2013. The Todai Robot Project. NII Today 46. (http://www.nii.ac.jp/userdata/results/pr_data/NII_Today/60_en/all.pdf).

Tandon, N.; Mishra, B. D.; Grus, J.; Yih, W.-t.; Bosselut, A.; and Clark, P. 2018. Reasoning about Actions and State Changes by Injecting Commonsense Knowledge. In EMNLP.

Trivedi, H.; Kwon, H.; Khot, T.; Sabharwal, A.; and Balasubramanian, N. 2019. Repurposing Entailment for Multi-Hop Question Answering Tasks. In NAACL.

Turney, P. D. 2006. Similarity of Semantic Relations. Computational Linguistics 32(3):379–416.

Turney, P. D. 2017. Leveraging Term Banks for Answering Complex Questions: A Case for Sparse Vectors. arXiv preprint arXiv:1704.03543.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. In ICLR.

Wang, W.; Yan, M.; and Wu, C. 2018. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. In ACL.

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv 1502.05698.

Wolfson, T.; Geva, M.; Gupta, A.; Gardner, M.; Goldberg, Y.; Deutch, D.; and Berant, J. 2020. Break It Down: A Question Understanding Benchmark. TACL.