Project Aristo: Towards Machines that Capture and Reason with Science Knowledge

Peter Clark
November 2019
The History of KCap

- KCap 2001-19
The History of KCap

- KCap 2001-19
- Banff Knowledge Acquisition Workshops: 1986-1999

How do we get knowledge into the machine in a usable form?
A Question for the field of Knowledge Capture

Structured Knowledge

Language (and other media)

relevant information

problem

K
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.
(D) The particles increase in volume.

Over the last decade, I began to think about a "Digital Aristotle", an easy-to-use, all-encompassing knowledge storehouse....to advance the field of AI.
Question Categories Not Covered

- **Diagrams**
  - [Diagram of Lifecycle Diagram]
  - [Diagram of Weather Map]
  - [Diagram of Graph]
  - [Diagram of Cell Structure]
  - [Diagram of Water Cycle]

- **Direct Answer Questions**
Progression on NY Regents 8th Grade (NDMC)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Progression on NY Regents 8th Grade (NDMC)

The graph shows the progression of baseline retrieval methods, tables plus rules, and a baseline without any retrieval methods. The baseline retrieval methods show an improvement from 36.4% in 2014 to 58.1% in 2015, followed by a slight increase to 63.1% in 2016. The tables plus rules method shows a similar pattern, with an increase from 25.0% in 2014 to 58.1% in 2015, followed by a slight increase to 63.1% in 2016. The random method remains constant at 0.0%.

(hidden test set, questions as written, NDMC, 5 years/119 qns)
The Allen Institute for Artificial Intelligence (AI2) is working to improve humanity through fundamental advances in artificial intelligence. One critical but challenging problem in AI is to demonstrate the ability to consistently understand and correctly answer general questions about the world.

The Aristo project at AI2 is focused on building such a system. One way Aristo "learns" is by extracting facts from various sources and processing them into a structured knowledge base. When taking an exam, questions are parsed and processed along with the knowledge base to generate a response.
Progression on NY Regents 8th Grade (NDMC)
Progression on NY Regents 8th Grade (NDMC)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Progression on NY Regents 8th Grade (NDMC)

The university of the state of New York
GRADE 8
INTERMEDIATE-LEVEL
SCIENCE TEST
WRITTEN TEST
JUNE 6, 2011

Student Name: ____________________________
School Name: ____________________________

Print your name and the name of your school on the lines above.
The questions on this test measure your knowledge and understanding of science.
The test has two parts. Both parts are contained in this test booklet.
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.
Part II consists of 25 open-ended questions. Write your answers to these questions in the space provided in this test booklet.
You may use a calculator to answer the questions on the test if needed.
You will have two hours to answer the questions on this test.

-- DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO. --

Copyright 2014
THE UNIVERSITY OF THE STATE OF NEW YORK
THE STATE EDUCATION DEPARTMENT
ALBANY, NEW YORK 12234

2014 36.4% baseline retrieval methods
2015 58.1% tables
2016 63.1% statistics + retrieval + reasoning (AAAI’16)
2017 72.2% (random)
2018 73.1%
2019 90.7%

Separate test on 3 latest exams (2017-2019): 93.3%

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
Aristo: an over-simplified overview

- An ensemble architecture
An ensemble architecture
Aristo: an over-simplified overview

- An ensemble architecture
1. Table Knowledge

In New York State, the longest period of daylight occurs during which month? (A) June (B) March  (C) December  (D) September
In New York State, the longest period of daylight occurs during which month? (A) June  
(B) March  (C) December  (D) September

- Daylengths in different months and locations?
- Solstices?
- Where is New York State?
- Which hemisphere is it in?
1. Table Knowledge: Aristo’s Tablestore

- ~120 tables, ~10-500 rows each
- Defined with respect to questions, study guides, syllabus
## IKE – Interactive Knowledge Extraction

### Query Expression Editor

<table>
<thead>
<tr>
<th>Add to Material-Conduct</th>
<th>Material</th>
<th>Energy</th>
<th>Count</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>material</td>
<td>heat</td>
<td>19</td>
<td>Less material dissipates heat slower and the sphere would maintain its temperature longer.</td>
</tr>
<tr>
<td></td>
<td>air</td>
<td>heat</td>
<td>15</td>
<td>Because air conducts heat much less readily than liquid does, less heat is transferred between the air and the absorber than in a</td>
</tr>
<tr>
<td></td>
<td>soil</td>
<td>heat</td>
<td>13</td>
<td>In addition, moist soil will conduct heat better than dry soil.</td>
</tr>
<tr>
<td></td>
<td>liquid</td>
<td>heat</td>
<td>12</td>
<td>But as a liquid absorbs heat energy, its molecules tend to vibrate more and more.</td>
</tr>
<tr>
<td></td>
<td>gas</td>
<td>heat</td>
<td>11</td>
<td>However, the gas conducts heat away so some additional power is wasted to heating the surroundings.</td>
</tr>
<tr>
<td></td>
<td>ice</td>
<td>heat</td>
<td>9</td>
<td>Sea ice formation releases heat during freezing conditions, and the melting of sea ice absorbs heat.</td>
</tr>
<tr>
<td></td>
<td>salt water</td>
<td>electricity</td>
<td>9</td>
<td>The idea here is that salt water conducts electricity, and the conductivity can</td>
</tr>
<tr>
<td></td>
<td>insulation</td>
<td>heat</td>
<td>8</td>
<td>Since insulation holds heat in, protection is provided by slowing down the loss of heat from young tree trunks, thus making them</td>
</tr>
<tr>
<td></td>
<td>cold water</td>
<td>heat</td>
<td>6</td>
<td>Answer: Cold water conducts heat away from the body 25 times faster than cold air because water has a much higher conductivity than air.</td>
</tr>
<tr>
<td></td>
<td>steam</td>
<td>heat</td>
<td>6</td>
<td>Conversion of the water to steam absorbs heat by reducing the oxygen content of the atmosphere and active burning should cease.</td>
</tr>
<tr>
<td></td>
<td>air</td>
<td>electricity</td>
<td>4</td>
<td>Even an isolating material such as air will conduct electricity during a thunderstorm since lightning bolts have such immense voltages.</td>
</tr>
<tr>
<td></td>
<td>salt water</td>
<td>heat</td>
<td>4</td>
<td>Solar ponds use the natural properties of salt water to collect and store heat</td>
</tr>
</tbody>
</table>

(AKBC’16)
1. Table Inference

In New York State, the longest period of daylight occurs during which month? (A) June (B) March (C) December (D) September
2. Table Inference

In New York State, the longest period of daylight occurs during which month? (A) June (B) March (C) December (D) September

<table>
<thead>
<tr>
<th>Subdivision</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York State</td>
<td>USA</td>
</tr>
<tr>
<td>California</td>
<td>USA</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>Brazil</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orbital Event</th>
<th>Day Duration</th>
<th>Night Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer Solstice</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>Winter Solstice</td>
<td>Short</td>
<td>Long</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>North</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

Semi-structured Knowledge
In **New York State**, the **longest period of daylight** occurs during which month? (A) June (B) March (C) December (D) September

#### Semi-structured Knowledge
In New York State, the longest period of daylight occurs during which month? (A) June (B) March (C) December (D) September

<table>
<thead>
<tr>
<th>Subdivision</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York State</td>
<td>USA</td>
</tr>
<tr>
<td>California</td>
<td>USA</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>Brazil</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orbital Event</th>
<th>Day Duration</th>
<th>Night Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer Solstice</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>Winter Solstice</td>
<td>Short</td>
<td>Long</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>North</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

Semi-structured Knowledge
In New York State, the **longest period of daylight** occurs during which month? (A) June  (B) March  (C) December  (D) September
## 2. Tuple Knowledge

![Diagram](image)

**Score** | **Tuple** | **Verbalization**
--- | --- | ---
... | ... | ... |
1.00 | most elephant isa mammal | // Elephant isa mammal. |
1.00 | most elephant isa pachyderm | // Elephant isa pachyderm. |
1.00 | most elephant require litre water | // Most elephants require litre water. |
1.00 | most elephant require water | // Most elephants require water. |
... | ... | ... |
0.92 | most elephant have curve spine | // Most elephants have curve spines. |
0.92 | most elephant need food | // Most elephants need food. |
... | ... | ... |
0.83 | most computer receive electric energy | // Most computers receive electric energy. |
0.67 | most computer solve problem | // Most computers solve problems. |
0.60 | most computer provide prediction | // Most computers provide predictions. |
... | ... | ... |
2. Tuple Inference

Which object in our solar system reflects light and is a satellite that orbits around one planet? (A) Moon (B) Earth (C) Mercury (D) Sun

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
Aristo: an over-simplified overview

- An ensemble architecture
Aristo: an over-simplified overview

- An ensemble architecture

![Diagram of Aristo ensemble architecture]

- Text
  - Web, Science

- Structured Representation
- Inference Solvers
- Language Models
- Retrieval & Statistics
- Combiner (Ensemble)

- Question

- Answer
What part of a plant needs sunlight to do its job? (A) leaf…
Where is the Knowledge Capture?

What part of a plant needs sunlight to do its job? (A) leaf
Where is the Knowledge Capture?

What part of a plant needs sunlight to do its job? (A) leaf

Wikipedia + BookCorpus (2.5B words + 11k books)
Where is the Knowledge Capture?

What part of a plant needs sunlight to do its job? (A) leaf

Wikipedia + BookCorpus (2.5B words + 11k books)

Curriculum Training

Allen Institute for Artificial Intelligence
Where is the Knowledge Capture?

What part of a plant needs sunlight to do its job? (A) leaf…

Wikipedia + BookCorpus (2.5B words + 11k books)

Curriculum Training

Aristo Corpus (2B words)
Exploiting Language Models

Earlier solvers + Multee → AristobERT

AristoBERT: 73.1
Exploiting Language Models

Earlier solvers + Multee = AristoBERT

![Bar chart showing comparison between Earlier solvers and Multee](chart.png)
Exploiting Language Models

Earlier solvers + Multee

AristoBERT
Exploiting Language Models

Earlier solvers + Multee

AristoBERT

73.1
70.8
79
67.2
Exploiting Language Models

Earlier solvers
+ Multee

AristoBERT

73.1
70.8
79
67.2
83.2
Exploiting Language Models

Earlier solvers

+ Multee

AristoBERT

Aristo

- Roberto

Earlier solvers

+ Multee

+ AristoBERT

73.1

70.8

79

67.2

83.2

87.4
Exploiting Language Models

Earlier solvers + Multee + AristoBERT
Exploiting Language Models

Earlier solvers + Multee

Aristo

BERT

AristoBERT

Earlier solvers + Multee + AristoBERT

ARISTO
Similarly on 12th grade NDMC:

- Random: 25.0%
- 2014: 40.6%
- 2019: 83.5%
## Individual Solver Performances

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Num Q</th>
<th>IR</th>
<th>PMI</th>
<th>ACME</th>
<th>TupInf</th>
<th>Multee</th>
<th>AristoBERT</th>
<th>AristoRoBERTa</th>
<th>ARISTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>109</td>
<td>64.45</td>
<td>66.28</td>
<td>67.89</td>
<td>63.53</td>
<td>69.72</td>
<td>86.24</td>
<td>88.07</td>
<td>89.91</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>119</td>
<td>66.60</td>
<td>69.12</td>
<td>67.65</td>
<td>61.41</td>
<td>68.91</td>
<td>86.55</td>
<td>88.24</td>
<td>91.60</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>632</td>
<td>41.22</td>
<td>46.95</td>
<td>41.57</td>
<td>35.35</td>
<td>56.01</td>
<td>75.47</td>
<td>82.28</td>
<td>83.54</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>2376</td>
<td>74.48</td>
<td>77.76</td>
<td>66.60</td>
<td>57.73</td>
<td>64.69</td>
<td>81.78</td>
<td>82.88</td>
<td>86.99</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>1172</td>
<td>n/a†</td>
<td>n/a†</td>
<td>20.44</td>
<td>23.73</td>
<td>37.36</td>
<td>57.59</td>
<td><strong>64.59</strong></td>
<td>64.33</td>
</tr>
</tbody>
</table>

Most of the heavy lifting....
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
1. Checking for annotation artifacts

The condition of the air outdoors at a certain time of day is known as

(A) friction
(B) light
(C) force
(D) weather

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>“Answer only” score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>38.53</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>37.82</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>47.94</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>36.17</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>35.92</td>
</tr>
<tr>
<td>All</td>
<td>37.11</td>
</tr>
</tbody>
</table>
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct]
The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct]

Retrain
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction  (E) joule
(B) light    (F) gradient [selected]
(C) force   (G) trench
(D) weather [correct, selected]

Retrain

Drop of (only) ≈ 10 points
City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) decreasing the cost of gasoline
(4) lowering the cost of bus and subway fares
City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) decreasing the cost of gasoline
(4) lowering the cost of bus and subway fares

Which of the following organs does a squirrel not have
(A) a brain
(B) gills
(C) a heart
(D) lungs
3. More than Pattern Matching?

City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) decreasing the cost of gasoline
(4) lowering the cost of bus and subway fares

Which of the following organs does a squirrel not have
(A) a brain
(B) gills
(C) a heart
(D) lungs
3. More than Pattern Matching?

2019 Report Card for

Subject  | Grade  | Teacher Comments
---|---|---
Negation  | A  | Nice work!
Conjunction
Polarity
World tracking
Factivity
Counting

Alan is small.  Alan is tall.  Bob is big.  Bob is tall.
Charlie is big.  Charlie is tall.  David is small.  David is short.

Which of the following is not tall? (A) Alan (B) Bob (C) Charlie (D) David [correct]
### 3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td><strong>Conjunction</strong></td>
<td></td>
<td>Nice work!</td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2019 Report Card for **Aristo**

- Grade: A
- Teacher Comments: Nice work!
- Percentage: 94%
Synthetic Conjunction Test

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

1 conjunct: 98%
2 conjuncts: 95%
3 conjuncts: 94.5%
4 conjuncts: 80%

+ 1 negation


Which of the following is old and red and light and big and not short? (A) Alan (B) Bob (C) Charlie (D) David...
3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td><strong>Conjunction</strong></td>
<td>B+</td>
<td><strong>Nice work!</strong></td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2019 Report Card for Aristo

- Subject
- Grade
- Teacher Comments

- Negation: A
- Conjunction: B+
- Polarity
- World tracking
- Factivity
- Counting

94%
80% - 98%
3. More than Pattern Matching?

2019 Report Card for Aristo

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
<td></td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td>Could ace this with more study!</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

94%
80% - 98%
67.1%

If Jim turns the thermostat down in his room while listening to music, what will happen to the speed of the sound waves in the room?

(A) they will speed up (B) they will slow down [correct] [correct]
## 3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
<th>Aristo</th>
<th>2019 Report Card for Aristo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
<td></td>
<td>94%</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
<td></td>
<td>80% -98%</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td>Could ace this with more study!</td>
<td></td>
<td>67.1%</td>
</tr>
<tr>
<td>World tracking</td>
<td>C</td>
<td></td>
<td></td>
<td>72.5%</td>
</tr>
</tbody>
</table>

**Context:** If someone travels for longer, they will travel further.

**Question:** John and Rita are going for a run. Rita gets tired and takes a break on the park bench. After twenty minutes in the park, who has run farther?

(A) John [correct] (B) Rita
### 3. More than Pattern Matching?

#### 2019 Report Card for **Aristo**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td>Could ace this with more study!</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td><strong>Factivity</strong></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 94%
- 80% -98%
- 67.1%
- 72.5%
- 66.5%

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

Subject | Grade | Teacher Comments | Aristo
---|---|---|---
Negation | A | Nice work! | 94%
Conjunction | B+ | Could ace this with more study! | 80% -98%
Polarity | D+ | 67.1%
World tracking | C | 72.5%
Factivity | D | 66.5%
Counting | F | 6%
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
4. Where is Aristo Failing?

What part of a plant needs sunlight to do its job? (A) leaf...
4. Where is Aristo failing?

- Case study on 30 failures:

  - Reading Comprehension (IR won’t help)
    - Good support for correct answer
      - 8 (26.7%)
    - Good support for incorrect answer
      - 4 (13.3%)
    - No good support
      - 17 (56.7%)
1. Good support for the correct answer (13%)

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

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

Distances between Earth and the stars are often measured in terms of light-years.
2. Good support for the incorrect answer (3%)

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

- I remember it had like a rubber ball in it, which would maybe float up…
- We played soccer with a giant rubber ball that floated like a balloon.
- Rubber toys floated on the water.
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 (D) their method of catching food

- Need question decomposition

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 (D) The particles increase in volume

- No good single supporting sentence
3. No good support for the correct answer (57%)

Which characteristic applies to animals in only one of these taxonomic groups: reptiles, mammals, birds, amphibians, or fishes? (A) have hair (B) lay eggs (C) have webbed feet (D) breathe with gills

- Boolean reasoning

Which geologic structure will most likely take the longest time to form? (A) a fault (B) a sinkhole (C) a river meander (D) a mountain range

- Cross-option comparative
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 (D) amount of growth

Which statement is an opinion? (A) Many plants are green. (B) Many plants are beautiful. (C) Plants require sunlight. (D) Plants can grow in different places.
About how long does it take for the Moon to complete one revolution around Earth? (A) 7 days (B) 30 days (C) 90 days (D) 365 days

- Because it takes the moon about **27.3 days** to complete one orbit around the Earth, the moon moves a little bit further around the Earth each day.
- It takes **27.3 days** for the moon to complete one revolution around the earth.
- The moon completes one revolution of the Earth in about **29.5 days**.
- The Moon completes one revolution around the Earth in **27.32166 days**.
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
1. Question Decomposition

What virus structure is similar in function to a cell membrane? (A) protein shell (B) internal protein...
What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...

What is the function of a cell membrane?

Surrounds and protects, gives structure, regulates material, ….

What part of the virus surrounds and protects it?

Protein shell, protein layer, …

- GapQA \((EMNLP'19)\)
- New dataset coming
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy
Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity. Electrical conductivity is the capacity of metal to conduct an electric current. Water without minerals will not conduct electricity.
2. Multihop Reasoning

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity. Electrical conductivity is the capacity of metal to conduct an electric current. Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working. A suit of armour is a historical type of personal body armour made from metal. Resisting arrest is a criminal charge, but civil suits can be filed.

Form Chains:
“suit of armor...made from metal” AND “…metal conduct electrical current”
=> “suit of armor conducts electricity”

“Resisting arrest...suits can be filed” AND “reciprocal of resistivity is conductivity”
=> “suit of armor conducts electricity”

Train system to recognize good chains
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.

Where is the sugar created? Light, water, CO2
### Paragraph

- **s1** Roots absorb water from soil.
- **s2** The water flows to the leaf.
- **s3** Light and CO2 enter leaf.
- **s4** Water, light, CO2 form sugar.

### State changes: \( \pi \)

<table>
<thead>
<tr>
<th></th>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Modeling World States

**Paragraph**

1. Roots absorb water from soil.
2. The water flows to the leaf.
3. Light and CO2 enter leaf.

**State changes: \( \pi \)**

<table>
<thead>
<tr>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>soil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>roots</td>
<td>s1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>s4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Modeling World States

**Paragraph**

s1 Roots absorb water from soil.

**State changes:** \( \pi \)

<table>
<thead>
<tr>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>soil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>roots</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

s2 The water flows to the leaf.

s3 Light and CO2 enter leaf.

s4 Water, light, CO2 form sugar.
3. Modeling World States

Paragraph

- s1 Roots absorb water from soil.
- s2 The water flows to the leaf.
- s3 Light and CO2 enter leaf.
- s4 Water, light, CO2 form sugar.

State changes: \( \pi \)

<table>
<thead>
<tr>
<th></th>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>soil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td>roots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td>leaf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Modeling World States

Paragraph

s1 Roots absorb water from soil.
s2 The water flows to the leaf.

s3 Light and CO2 enter leaf.

s4 Water, light, CO2 form sugar.

State changes: $\pi$

<table>
<thead>
<tr>
<th></th>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>soil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2</td>
<td>roots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s3</td>
<td>leaf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s4</td>
<td>leaf</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Modeling World States

**Paragraph**

1. Roots absorb water from soil.
2. The water flows to the leaf.
3. Light and CO2 enter leaf.

**State changes:** \( \pi \)

<table>
<thead>
<tr>
<th></th>
<th>water</th>
<th>light</th>
<th>CO2</th>
<th>sugar</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>soil</td>
<td>sun</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>s2</td>
<td>roots</td>
<td>sun</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>s3</td>
<td>leaf</td>
<td>sun</td>
<td>?</td>
<td>-</td>
</tr>
<tr>
<td>s4</td>
<td>leaf</td>
<td>leaf</td>
<td>leaf</td>
<td>-</td>
</tr>
</tbody>
</table>
4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes

Why?

Because
• pennies are made of metal
• metals are magnetic

Actually:
Not all metals are magnetic.
Copper is not magnetic.

Try again!

No – because:
• pennies are made of copper
• copper is not magnetic
A Question for the field of Knowledge Capture

Structured Knowledge

Language (and other media)

problem

relevant information
A Question for the field of Knowledge Capture

relevant information

problem

Structured Knowledge

Language (and other media)
Summary

- Surprising success!
  - LMs: Structure not essential for many tasks
  - >> “just pattern matching”
- BUT:
  - falls short with numerous types of questions
  - many other AI aspects missing

What do we need going forward?
- Structured reasoning and knowledge capture but with more language-like representations

Thank you!