In Pursuit of the Holy Grail of Natural Language Understanding:
Past, Present, and Future

September 2019

Daniel Khashabi
My Background

- Recently finished PhD (UPenn)
- Been working with Allen Institute for AI (AI2)
  - On-and-off since 2015, and full-time since this August

- Research theme:
  - Artificial Intelligence, through the lens of **natural language** understanding

- Not specified:
  - What **solution** we use to achieve this goal

- Not a CP person!
• Founded in 2014 by Paul Allen (Microsoft co-founder)
  • Non-profit research organization
  • Mission: contribute to humanity through high-impact AI research and engineering
Project Aristo (2014-2019)

- **Vision: The Knowledgeable Machine**
  - Large volumes of general and scientific knowledge, stored in a "computable" form, supporting reasoning and explanation

- **Measurable goal:**
  - Pass elementary-school **science exams** as written
    - Currently 4th grade and 8th grade exams

- Credit goes to dozens of researchers who have contributed to this project.
Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the (A) cell membrane. (B) cytoplasm. (C) cell wall. (D) ribosome.
Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the
(A) cell membrane.
(B) cytoplasm.
(C) **cell wall**.
(D) ribosome.

Plants use **cellulose** for their cell walls.
The cell structure that makes a plant cell more rigid than an animal cell is the (A) cell membrane. (B) cytoplasm. (C) cell wall. (D) ribosome. 

Plants use cellulose for their cell walls. All fibers are assembled of chains of cellulose molecules, arranged as a rigid structure.
Example Question: Reasoning by Chaining

The cell structure that makes a plant cell more rigid than an animal cell is the
(A) cell membrane.
(B) cytoplasm.
(C) cell wall.
(D) ribosome.

Plants use **cellulose** for their cell walls.

All fibers are assembled of chains of cellulose molecules, arranged as a rigid structure.

Plants use cellulose for their cell walls.
Aristo: an over-simplified overview

- An ensemble architecture
  - To deal with questions with a variety of difficulty

![Diagram of Aristo architecture]
Road map

I. Aristo Inference Solvers

II. Beyond inference Modules
Aristo Inference Modules

Khashabi et al, AAAI-2018
Khot et al, ACL-2017
Khashabi et al, IJCAI-2016
Aristo Inference Modules

- General structure of **inference frameworks**
  - Text & questions converted to an intermediate representation
  - A “probabilistic” inference for question-answering

- Three key scientific questions:
  1. How do we identify and represent the **meaning** of text?
  2. What mechanisms support **robust reasoning** with incomplete and incorrect knowledge?
Knowledge in Tabular Representation

Text

Knowledge Base
(library of ontologies and domain models)

Tractable, probabilistic Inference

One giant database of science knowledge

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>northern</td>
<td>summer solstice</td>
<td>Jun</td>
</tr>
<tr>
<td>northern</td>
<td>winter solstice</td>
<td>Dec</td>
</tr>
<tr>
<td>northern</td>
<td>autumn equinox</td>
<td>Sep</td>
</tr>
<tr>
<td>southern</td>
<td>summer solstice</td>
<td>Dec</td>
</tr>
<tr>
<td>southern</td>
<td>autumn equinox</td>
<td>Mar</td>
</tr>
</tbody>
</table>

Energy, Forces, Adaptation, Phase Transition, Organ Function, Tools, Units, Evolution, ...

Simple structure, flexible content
- Can acquire knowledge in automated and semi-automated ways
- No Cyc-like claims of “completeness” or “adequacy”  

[Dalvi et al, 2016]
Q: In New York State, the longest period of daylight occurs during which month?

<table>
<thead>
<tr>
<th>Cities, States, Countries</th>
<th>Orbital Events:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential Link:</td>
<td>Geographical properties &amp; Timing</td>
</tr>
<tr>
<td>Regions and Hemispheres</td>
<td></td>
</tr>
</tbody>
</table>

- (A) December
- (B) June
- (C) March
- (D) September
TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Q: In New York State, the longest period of daylight occurs during which month?

<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**
Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Q: In **New York State**, the **longest period of daylight** occurs during which **month**?

<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>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>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>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>

**Abductive reasoning**

[Peirce, 1883]
TableILP Solver: Overview

A discrete **optimization** approach to QA for multiple-choice questions

**ILP model builder**

**M(T,Q,A)**

**ILP engine**

**M(G,Q,A)**

\[
\begin{align*}
\max \sum_i c_i x_i \\
\forall x \in \mathbb{N} \cup \{0\}
\end{align*}
\]

\[
\begin{align*}
\sum_i a_{1i} x_i & \leq b_1 \\
\cdots \\
\sum_i a_{ki} x_i & \leq b_k
\end{align*}
\]

Optimization using Integer Linear Program (ILP) formalism
**Goal:** Design ILP objective function, s.t. maximizing it subject to the constraints yields a “desirable” support graph.

**Natural Language (Question, …)** ➔ **Conceptual def of: Constraints, Objective** ➔ **ILP CP SAT …** ➔ **Solution**

- **What we study ✓**
- **Not our focus ✗**
Goal: Design ILP objective function, s.t. maximizing it subject to the constraints yields a “desirable” support graph

Not so straightforward!

- Many possible “proof structures”
- Imperfect lexical “similarity” blackbox
- Partial or missing knowledge
- Question logic (negation, conjunction, comparison)
- Scalability of ILP solvers
- …
ILP Model: Some Details

\[
\begin{align*}
\text{max} & \quad \sum_i c_i x_i \\
\text{subject to} & \quad \sum_i a_{1i} x_i \leq b_1 \\
& \quad \sum_i a_{ki} x_i \leq b_k \\
\forall x_i & \in \mathbb{N} \cup \{0\}
\end{align*}
\]

Q: In New York State, the **longest** period of daylight occurs during which **month**?

- (A) December
- (B) June
- (C) March
- (D) September

<table>
<thead>
<tr>
<th>Subdivision</th>
<th>Country</th>
<th>Orbital Event</th>
<th>Day Duration</th>
<th>Night Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York State</td>
<td>USA</td>
<td>Summer Solstice</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>California</td>
<td>USA</td>
<td>Winter Solstice</td>
<td>Short</td>
<td>Long</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>Brazil</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Winter Solstice</td>
<td>June</td>
<td></td>
</tr>
</tbody>
</table>
Variables define the space of “support graphs”:

- Each variable corresponds to a node or edge.
- \( x = 1 \) iff nodes / edges are part of the semantic graph.

\[
\max \sum_i c_i x_i \\
\sum_i a_{1i} x_i \leq b_1 \\
\ldots \\
\sum_i a_{ki} x_i \leq b_k
\]

\[\forall x_i \in \mathbb{N} \cup \{0\}\]

Support Graph

- Q: In New York State, the longest period of daylight occurs during which month?

Semi-structured Knowledge
Objective Function:

“better” support graphs = higher objective value

- Reward good behavior:
  - High lexical match links, nearby alignments, column header match, WH-terms (“which of energy …”), etc.

- Penalize spurious overuse of frequently occurring terms

\[
\max \sum_i c_i x_i \quad \forall x_i \in \mathbb{N} \cup \{0\}
\]

\[
\sum_i a_{1i} x_i \leq b_1 \\
\sum_i a_{ki} x_i \leq b_k
\]
Constraints:

- ~50 high-level constraints
- Scalability + consider only meaningful support graphs

**Structural Constraints:**
Many possible Proof Structures:
  - Basic Lookup, Parallel Evidence, Evidence Chaining, Semantic Relation Matching

**Semantic Constraints:**
Meaningful proof structures
  - Connectedness, question coverage, appropriate table use, etc.

\[
\max \sum_{i} c_i x_i \\
\forall x_i \in \mathbb{N}\cup\{0\}
\begin{align*}
\sum_{i} a_{1i} x_i & \leq b_1 \\
\sum_{i} a_{ki} x_i & \leq b_k
\end{align*}
\]
Some Empirical Results

Ensemble performs 8-10% higher than IR baselines

More details in: K et al, IJCAI-2016.
Beyond Tables

- Issues with tabular representation:
  - Hard to extract; the schemas could be limiting

- Alternative representation:
  - Rich representation
  - Easy to automatically extract

- Idea: **Reason over (multiple) semantic abstractions of text**
  - Use off-the-shelf, pre-trained NLP modules for automatic extraction of representations

- Same ILP-based reasoning formalism as before
NLP field has developed tools to extract and represent many interesting phenomena in language:

- **Example 1:** Co-reference

... *Polar bears, saved from the bitter cold by their thick fur coats, are among the animals in danger of extinction...*
Mapping Text to Semantic Representations

- NLP field has developed tools to extract and represent many interesting phenomena in language:

- **Example 3:** events described by verbs

  ... Polar bears, **saved from the bitter cold by their thick fur coats,** are among the animals in danger of extinction ...

**Verb Semantic Roles** [Punyakanok et al. 2008]
Inference Solvers, Beyond Tables

- Extend model to consume various semantic representations.
  - *Trickier than it appears:* Can no longer rely on tables as a coherent collection of domain-relevant pieces of knowledge
  - Must revisit tuple selection, chaining, …

Khashabi et al, AAAI-2018
Khot et al, ACL-2017
Some Empirical Results

Top scoring systems in their own time.

More details in the related papers

Ket al, AAAI-2018
Khot et al, ACL-2017
interim Summary

- The importance of the ability to chain information

- Structured, Multi-Step Reasoning:
  - Use science knowledge in small, reusable, swappable pieces

- State-of-the-art results on science benchmarks

- Benefits of the approach:
  - Principled approach
  - Explainable answers
Machine Reasoning [in Aristo]:
Today and Future
Progression on NY Regents 8th Grade (NDMC)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Aristo aces 8th grade (non-diagram multiple choice) >90%

(hidden test set, questions as written, NDMC, 5 years/119 qns)
A Breakthrough for A.I. Technology: Passing an 8th-Grade Science Test

By Cade Metz
Sept. 4, 2019

SAN FRANCISCO — Four computer scientists competed in a contest in June to see if they could pass an eighth-grade science test. They all flunked. Even the best artificial intelligence systems fell far short of the raw and logic skills that students are expected to master before they graduate from high school.

But on Wednesday, the Allen Institute, a prominent lab in Seattle, unveiled a new system that passed the test with room to spare. It correctly answered more than 90 percent of the questions on an eighth-grade science test and more than 80 percent on a 12th-grade exam.

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 in 2016, the best AI system achieved merely 59.3% on an 8th Grade science exam challenge (Schoenick et al., 2016).

What constraints are there on the interaction? What guidelines are provided to the judges? Second, recent Turing Test competitions have shown that, in certain formulations, the test itself is gameable; that is, people can be fooled by systems that simply retrieve sentences and make no claim of being intelligent (Aron, 2011; BBC, 2014). John Markoff of The New York Times wrote that the Turing Test is more like a magic trick than a test of intelligence, though it is an interesting question whether it can be improved upon.
Neural Network solvers for natural language

- Big gains are achieved in the past 1-2 years
- Not just in Aristo: they have been pretty effective in a wide range of NLP problems

Many works:
[Peters et al. 18, Devlin et al.18; Radford et al.18, ...]
My view:
- An engineering revolution
- Concepts and ideas were out there (for decades)

Enabled by:
- **Big compute:**
  - Having access to many GPUs/TPUs is a must now

- **Massive pre-training:**
  - Billions of documents are used to tune them

- **Big supervision:**
  - Availability of more annotated data
Neural Net resurgence
What’s next? What is missing?

- Neural Nets have nearly “solved” the science challenges!
  - Do we understand why?

- In many cases, we are not able to explain their decisions.
  - As a result, we can’t assess the extent of their reliability.

- Additionally, extending them is not trivial (largely an open problem)
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

So maybe it does actually work? 🤔
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

Hmm 🤔 ... but how should we fix this??
The Aristo experiment:

- Working hypothesis: Real-world language and reasoning capabilities can be assessed via “well-designed” QA tasks.

First systems for science QA effective at multi-hop inference

- Can operate with semi-formal knowledge bases
- Can fruitfully exploit noisier & less structured representations
- Complementary to other methods

Deep learning is playing an important role, but is not everything (yet)

Interacting intelligently with humans remains challenging!

- Requires rich linguistic understanding, common sense, domain knowledge, situational awareness, conversational memory …
Questions?

Hiring talented interns, young investigators, researchers, engineers, and more!
Here are some options I found nearby:

Pattaya
Thai - 800 feet
⭐⭐⭐⭐⭐ (278) on Yelp · $$

Bobby's Burger Palace
American (Traditional) - 450 feet
⭐⭐⭐⭐⭐ (756) on Yelp · $$

POD
Asian Fusion - 0.2 miles
⭐⭐⭐⭐⭐ (731) on Yelp · $$$

Han Dynasty
Szechuan - 0.2 miles
⭐⭐⭐⭐⭐ (658) on Yelp · $$

Show me some restaurants nearby.

I don’t like crowds.

I am allergic to peanuts.
What needs to happen?

1. Define representation of functions
2. Manually author and automatically learn extraction rules
3. Attack challenges
   - Linguistic variability
     - “synthesize protein”, “protein synthesis”, “synthesizes large numbers of proteins”, “helps to synthesize proteins”, …
   - Multiword expressions
     - “cell membrane” = “membrane”, but “amino acid” ≠ “acid”
   - Ambiguous patterns
     - “cell membrane”, “cell division”, “cell biologist”
   - Non-functional actions
     - exist, differ, attempt, do, remain, appear, occur, play, show, …
   - Peripheral activities
     - nucleus + divide, ribosome + move
Aristo in Context

**Watson**

Knowledge Base (library of ontologies and domain models)

Natural Language Processing

Candidate answer extraction

Candidate answer phrases

Answer scoring + evidence combination

Ranked answers

**Aristo**

Knowledge Base (monolithic, fixed ontology)

Heavy-duty deductive inference

Knowledge Base

Candidate answer extraction

Candidate answer phrases

Answer scoring + evidence combination

Ranked answers

**Cyc**

Knowledge Engineering

Natural Language Processing

Candidate answer extraction

Candidate answer phrases

Answer scoring + evidence combination

Ranked answers

**Knowledge Engineers**

Knowledge Base

Candidate answer extraction

Candidate answer phrases

Answer scoring + evidence combination

Ranked answers

**Text**

Tractable, probabilistic Inference

**Text**

95%

34%

Watson

Aristo

Cyc

Knowledge Engineers
Example

**QUESTION**

Water freezing is an example of a (A) liquid turning into a solid (B) solid turning into a liquid (C) gas turning into a solid

**NEW:** object(freeze, water) -> d-change(water, -, liquid, solid)?

**TEXT**

Freezing involves changing water from its liquid state to its solid state by the removal of heat.

freeze -> d-change(water, -, liquid state, solid state).

Water freezing is an example of a (A) liquid turning into a solid (B) solid turning into a liquid (C) gas turning into a solid.

Freezing involves changing water from its liquid state to its solid state by the removal of heat.

NEW: \( \text{object(freeze, water)} \rightarrow \text{d-change(water, -, liquid, solid)?} \)

explicit representation of discrete change
What do earthquakes tell scientists about the history of the planet?

(A) Earth's climate is constantly changing.
(B) The continents of Earth are continually moving.
(C) Dinosaurs became extinct about 65 million years ago.
(D) The oceans are much deeper today than millions of years ago.
**Question:** What do earthquakes tell scientists about the history of the planet?

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

**Confidence:** 45.42%

as computed from these reasoners:

- Information Retrieval: 87.59%
- Table Reasoning: 6.68%
- Topic Matching: 65.20%
- Tuple Reasoning: 4.16%

**Justification sentence (from Information Retrieval):** The paleontological records of the history of life on this planet show that the giant dinosaurs and many other animal and plant groups became extinct about 65 million years ago.
City administrators can encourage energy conservation by

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

*Aristo is not sure about this one...*

**Best guess:** (C) decreasing the cost of gasoline

**Confidence:** 35.55%

as computed from these reasoners:

- *Information Retrieval:* 14.95%
- *Topic Matching:* 70.15%
- *Tuple Reasoning:* 69.43%

**Justification sentence (from Information Retrieval):** 1970s Programs and educational materials are created to encourage gasoline and energy conservation.
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.
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.

Retrieved knowledge:

As the heat of a particle increases, the particles move faster.
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.

**Retrieved knowledge**

As the heat of a particle increases, the particles move faster.

**Identified spans for qualitative property:**
- particles (0.77)
- particle (0.65)
- heat (0.62)

**Identified spans for qualitative direction:**
- more rapidly (0.93)
- faster (0.93)
- increases (0.93)
Example Question: Simple Lookup

How many chromosomes does the human body cell contain?

(A) 23
(B) 32
(C) 46
(D) 64
Which features are the best evidence that glaciers once covered an area?

(A) wide riverbeds
(B) U-shaped valleys
(C) groundwater springs
(D) underground caves

Aristo KB facts:

(glacier, causes, glacial erosion) &
(glacial erosion, causes, valley)

→ (glacier, causes, valley)
Handling Lexical Variability

Knowledge is expressed in a **variety of linguistic forms**
- Simple textual variation confuses even the best solvers
- No single knowledge representation (e.g., Open IE tuples) suffices

- **Idea:** *Reason over (multiple) semantic abstractions of text*
  - Use off-the-shelf, pre-trained NLP modules
  - Multiple views for a more complete semantic understanding
    - SRL frames (verbs, prepositions, comma), dependency parse, coreference sets, lexical similarity links, raw text sequence

- Unified representation as a family of **Semantic Graphs**
  - PredArg graphs, trees, clusters, sequences
  - Connected via textual similarity links

- Same ILP-based reasoning formalism as before