COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Semi-Structured Reasoning for Answering Science Questions

Daniel Khashabi, Dan Roth (UIUC)

Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni (Allen Institute for Artificial Intelligence)

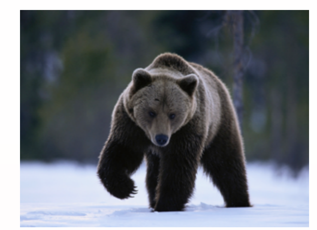
Overview

Challenge: Build an AI system that demonstrates human-like intelligence by passing standardized science exams as written; examples:

Intermediate Goal: Elementary Science: A simple embodiment of this challenge, requires question-answering significantly beyond retrieval techniques

Approach: A discrete optimization approach to QA for multiple-choice questions

Results: State of the art performance on 4th grade science (NY Regents exam)

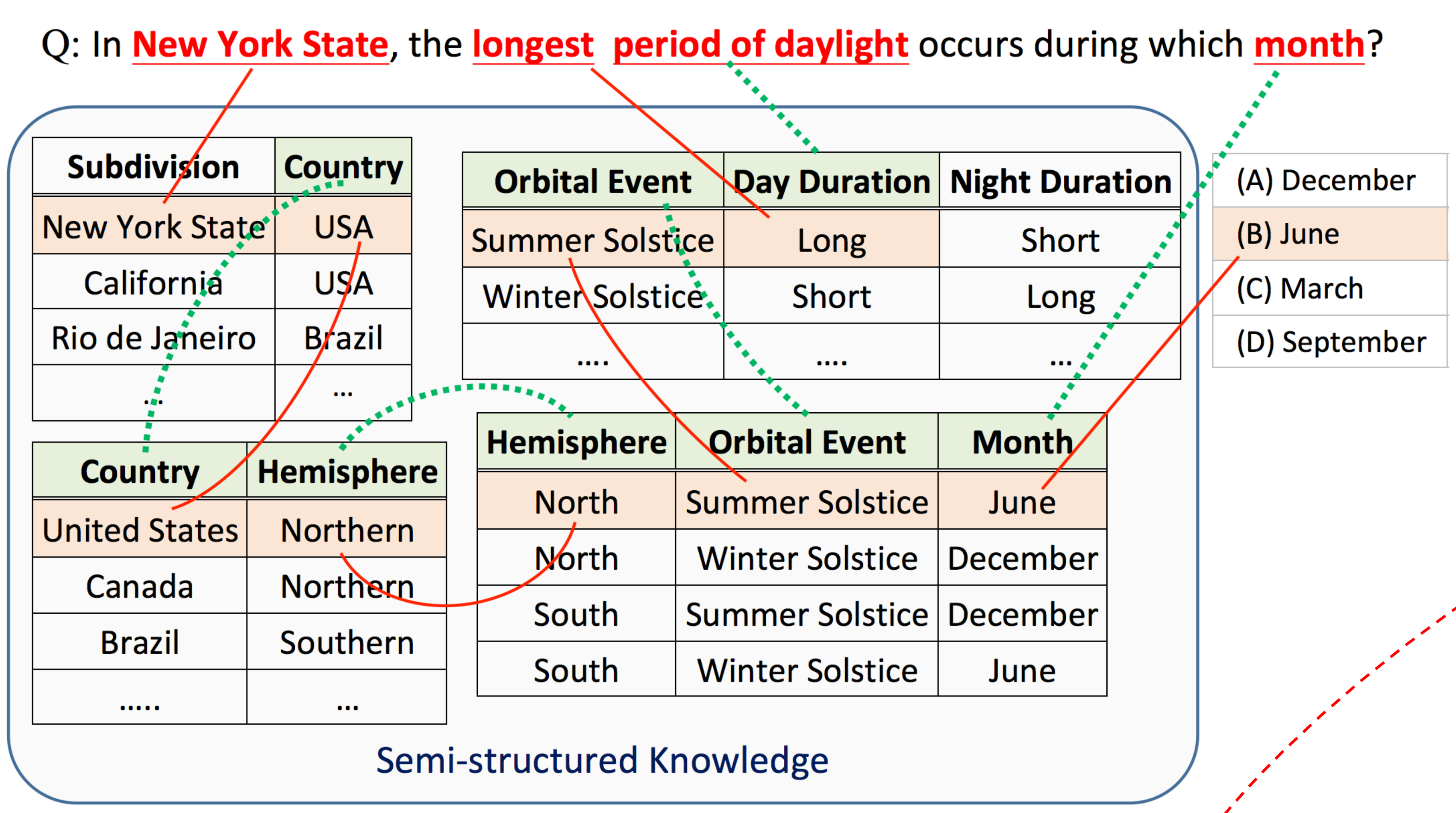


Which physical structure would best help a bear to **survive a winter** in New York State? (A) big ears (B) black nose (C) **thick fur** (D) brown eyes

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

Main Idea

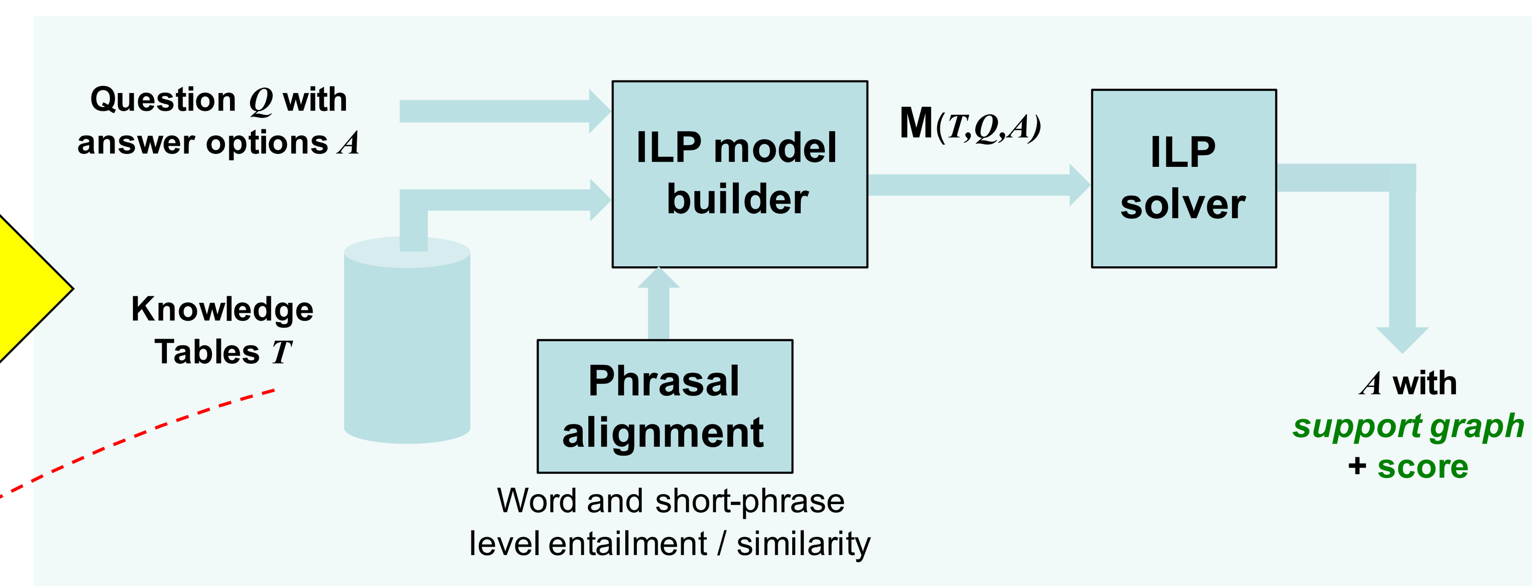
Search for the best Support Graph connecting the Question to an Answer through Tables.



Architecture

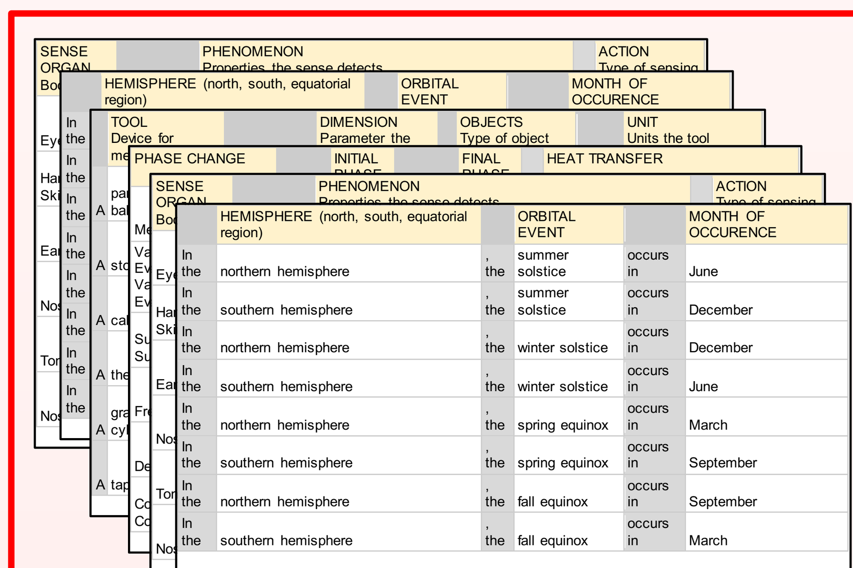
A discrete **optimization** (Integer Linear Program) approach to QA for multiple-choice questions

Given as input a question, candidates answers, and a generic set of tables, and generates a constrained optimization problem



$$M(T,Q,A) \rightarrow \max \sum_i c_i x_i \quad \forall x_i \in \mathbb{N}\cup\{0\} \quad \begin{cases} \sum_i a_{1i} x_i \le b_1 \\ \dots \\ \sum_i a_{ki} x_i \le b_k \end{cases}$$

- 4^{th} **Grade NY Regents Science Exam**
  - 108 training set
  - 129 questions in completely unseen Test set
- **Baselines:**
  - PMI Solver: Statistical correlation using pointwise mutual info (280 GB of text)
  - IR Solver: Information Retrieval using Lucene search (280GB of text)
    - IR Solver (table): using only Tablestore
  - MLN Solver: using rules from 80K sentences



69 tables, ~7.6K rows

Tablestore

- **Variables** define the space of "support graphs" connecting Q, A, T
  - Which nodes + edges between lexical units are active?
- **Objective** "better" support graphs = higher objective value
  - Reward active units, high lexical match links, column header match, …
  - WH-term boost (which form of energy), science-term boost (evaporation)
  - Penalize spurious overuse of frequently occurring terms

Objective function

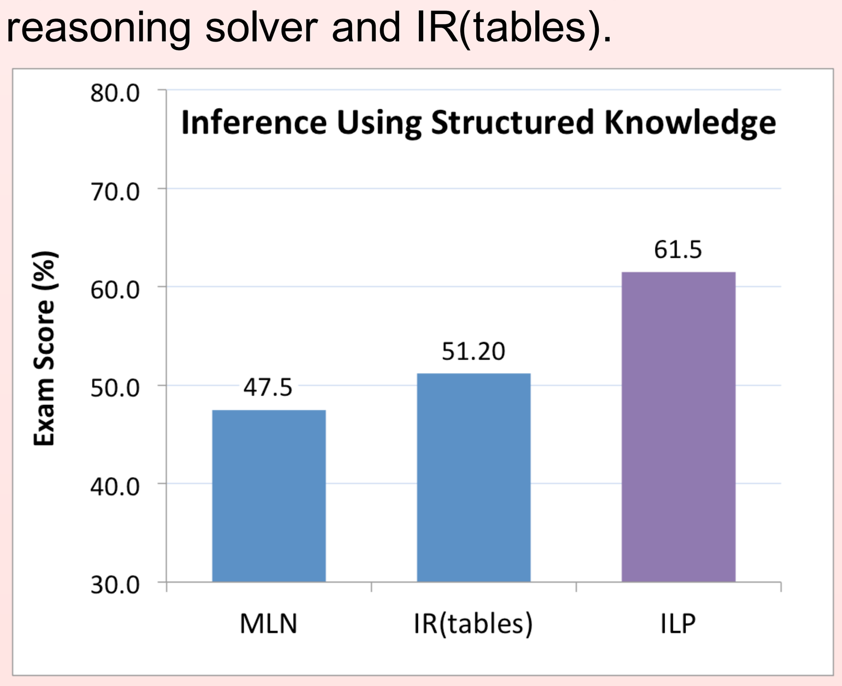
- **Structural Constraints**
  - Meaningful proof structures
    - connectedness, question coverage, appropriate table use
    - parallel evidence => identical multi-row activity signature
  - Simplicity appropriate for 4^{th} / 8^{th} grade
- **Semantic Constraints**
  - Chaining => table joins between semantically similar column pairs
  - Relation matching (ruler measures length, change from water to liquid)

Constraints

Evaluation

Exploiting Structured Knowledge

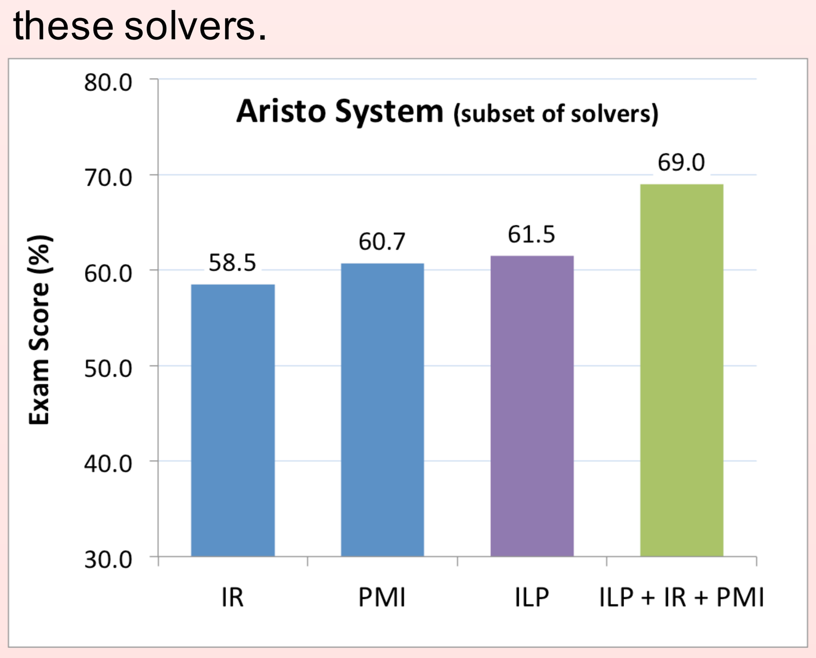
We compare the accuracy of our approach against the previous structured (MLN-based) reasoning solver and IR(tables).



1. TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources (6 years of exams; 95% C.I. = 9%)

Complementary Strengths

TableILP and IR-based methods clearly approach QA very differently. This analysis highlights the complementary strengths of these solvers.



2. Performance: % correct on 6 years of unseen questions (129 questions). The solvers ensemble performs 8-10% higher than IR baselines.

Assessing Brittleness: Question Perturbation

How robust are approaches to simple question perturbations *that would typically make the question easier for a human*? We consider a simple, automated way to perturb each 4-way multiple-choice question using Bing:

In New York State, the longest period of daylight occurs during which month?
(A) eastern (B) June (C) history (D) years

Solver	Original Score (%)	% Drop with Perturbation absolute	% Drop with Perturbation relative
IR	70.7	13.8	19.5
PMI	73.6	24.4	33.2
TableILP	85.0	**10.5**	**12.3**

3. On 1080 perturbed question of the regents train, TableILP has the smallest drop among the solvers.

ILP complexity, scalability

The table below summarizes various ILP and support graph statistics for TableILP, averaged across all test question

Category	Quantity	Average
ILP complexity	#variables	1043.8
	#constraints	4417.8
	#LP iterations	1348.9
Knowledge use	#rows	2.3
	#tables	1.3
Timing stats	model creation	1.9 sec
	solving the ILP	2.1 sec

4. Speed: 4 sec per question, reasoning over 140 rows across 7 tables. Contrast: 17 sec for MLN using only 1 rule per answer option!

Authors acknowledge the support from AI2 and DARPA (under the DEFT program).