Reasoning-driven Question Answering

Daniel Khashabi, Dan Roth
(Cognitive Computation Group, UPenn)

Tushar Khot, Ashish Sabharwal
(Allen Institute for Artificial Intelligence)
“A student leaves a bucket of water outside on a warm, sunny day. The water disappeared after a few hours. What phenomena describes this?”

“Evaporation”
What we know for sure

But there are certain things that we know for sure.

A “good” solution has to have:

- knowledge
- knowledge representation
- “easy” way of accessing the knowledge
- a decision making mechanism

This does not entail isolation/independence of these elements.

Natural Input

“A student leaves a bucket of water outside on a warm, sunny day. The water disappeared after a few hours. What phenomena describes this?”

World /background knowledge

∀x, y[brother(x, y) ⇒ sibling(x, y)]

Intermediate Input

Computer Brain

Intermediate Output

Natural Output

“Evaporation”
QA is everywhere

- One of the oldest problems in AI
- Remarkable features of QA

QA systems are still far from exhibiting human-like intelligence, even in relatively simple ways (vs. human-level)
General Problem Solver

(Simon & Newell, 1956)

Goal: Program for proving theorems!

Necessity: Representation with symbols!

Hypothesis (physical symbol system hypothesis):
“A physical symbol system has the necessary and sufficient means for general intelligent action.”

Reasoning: Problem solving as Search!
Programs with Commonsense

(John McCarthy, 1959)

Formalize world in **logical** form!

**Example:**
“My desk is at home” → \( \text{at}(I, \text{desk}) \)
“Desk is at home” → \( \text{at}(	ext{desk}, \text{home}) \)

**Hypothesis:** Commonsense knowledge can be formalized with logic.

Do **reasoning** on formal premises!

**Example Contd.:**
\[
\forall x \forall y \forall z \quad \text{at}(x,y), \text{at}(y,z) \Rightarrow \text{at}(x,z) \\
\therefore \text{at}(I, \text{home})
\]

**Hypothesis:** Commonsense problems are solved by logical reasoning
What they missed ….

- The difficulty of mapping from nature (including natural language) to symbols

One cannot simply map natural language to a representation that gives rise to reasoning

“Chicago”

Meaning

Variability

Language

Ambiguity
What they missed ....

- Reasoning is often studied in a very narrow sense.

Reasoning has many (infinite?) forms.

- One can think of it as a $n$ dimensional space
- Examples typically span multiple reasoning aspects.
How do you define “reasoning”? It’s a convoluted subject and hard to define.

reasoning in British (ˈrɪznɪŋ) noun
1. the act or process of drawing conclusions from facts, evidence, etc
2. the arguments, proofs, etc, so adduced

Collins English Dictionary. Copyright © HarperCollins Publishers
An Example for NLU …

AN EXAMPLE FOR NATURAL LANGUAGE UNDERSTANDING AND THE AI PROBLEMS IT RAISES

John McCarthy
Computer Science Department
Stanford University
Stanford, CA 94305
jmc@cs.stanford.edu
http://www-formal.stanford.edu/jmc/
1976
An Example for NLU …

(John McCarthy, 1976)

A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday in his downtown Brooklyn store by two robbers while a third attempted to crush him with the elevator car because they were dissatisfied with the $1,200 they had forced him to give them. The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.

... (McCarthy, 1990, p. 70)

- Predictions: Who had the money at the end?
- Yes/No question: Did Mr. Hug want to be crushed?
- What if question: What would have happened if Mr. Hug had not flattened himself at the bottom of the pit?

Not mentioned directly. It is implied.

Requires common sense + changing the knowledge according to the alternative scenario
Where are we?

How much “reasoning” do models do?

Neural networks
- sophisticated word-matching ~ paraphrasing

PMI-solver
- (Clark et al, 2016)

IR-solver (Lucene)
- (Clark et al, 2016)

MLNSolver
- (Khot et al, 2015)

Reasoning Axis

More Reasoning

How much do system decouple their language understanding from QA?

Neural networks
- sophisticated word-matching ~ paraphrasing

MLNSolver
- (Khot et al, 2015)

PMI-solver
- (Clark et al, 2016)

language Understanding decompoled from QA

IR-solver (Lucene)
- (Clark et al, 2016)

More decoupled
The Roadmap

- **Motivation**
  - Reasoning: Past and now

- **TableILP:** science QA with tables a knowledge
  1. Motivating example
  2. Tables as Knowledge
  3. Tables + ILP = TableILP solver

- **SemanticILP:** reasoning with layers of representation
  1. Motivating example
  2. Text + off-the-shelf annotators as knowledge
  3. Layers of semantic representations + ILP = SemanticILP solver

- **Summary**
Standardized Tests as an AI Challenge

Build AI systems that demonstrate human-like intelligence by passing standardized science exams as written.

Many challenges: broad knowledge (general and scientific), question interpretation, reasoning at the right level of granularity, …

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

**Premise:** A system that “understands” this phenomenon can correctly answer many variations!
Google is missing this behavior

what's the biggest airport in moscow

Domodedovo International Airport
Russia's busiest airports by passenger traffic in 2016

what's the smallest airport in moscow

List of airports in Russia - Wikipedia
List of airports in Russia (Russian Federation), sorted by location. There are 270 airports ... Moscow, UUDD, DME, Domodedovo International Airport · Moscow · Khodinka Airport · Moscow, UUMO, OSF, Ostafyevo International Airport · Moscow, UUEE ...

List of the busiest airports in Russia - Wikipedia
https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_Russia
This is a list of the busiest airports in Russia, using data from the Federal Air Transport Agency. ..... 7.4%, Steady: 2, Sheremetyevo International Airport · Moscow

Sheremetyevo International Airport - Wikipedia
https://en.wikipedia.org/wiki/Sheremetyevo_International_Airport
Sheremetyevo International Airport (IATA: SVO, ICAO: UUEE) is an international airport located .... The Moscow Oblast government has reserved adjacent land for a future third runway. ..... Tools. What links here · Related changes · Upload file · Special pages · Permanent link · Page information · Wikidata item · Cite this page ...
Semi-Structured Inference

- Structured, Multi-Step Reasoning
  - science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations

**How can we achieve this?**
Knowledge as Relational Tables

Unstructured

e.g., free form text from books, web

easy to acquire, difficult to reason with

Structured

e.g., probabilistic first-order logic rules, ontologies

“easy” to reason with, difficult to acquire

Relational Tables with free form text

collections of recurring, related, science concepts

<table>
<thead>
<tr>
<th>Country</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>north hemisphere</td>
</tr>
<tr>
<td>USA</td>
<td>north hemisphere</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>south hemisphere</td>
</tr>
<tr>
<td>Zambia</td>
<td>south hemisphere</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>northern</td>
<td>summer solstice</td>
<td>Jun</td>
</tr>
<tr>
<td>northern</td>
<td>winter solstice</td>
<td>Dec</td>
</tr>
<tr>
<td>...</td>
<td>autumn equinox</td>
<td>Sep</td>
</tr>
<tr>
<td>southern</td>
<td>summer solstice</td>
<td>Dec</td>
</tr>
<tr>
<td>...</td>
<td>autumn equinox</td>
<td>Mar</td>
</tr>
</tbody>
</table>

Simple structure, flexible content
- Can acquire knowledge in automated and semi-automated ways

Available at allenai.org
### 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>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 (i.e., best explanation)

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

<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></td>
<td></td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

Link this information to identify the best supported answer!
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>
</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>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**
TableILP Solver: Overview

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

- for each given question and candidate answers, we automatically generate a corresponding ILP objective and a set of constraints.

\[
M(T,Q,A) = \max \sum_i c_i x_i \quad \forall x_i \in \mathbb{N} \cup \{0\}, \quad \begin{cases} 
\sum_i a_{1i} x_i \leq b_1 \\
\vdots \\
\sum_i a_{ki} x_i \leq b_k 
\end{cases}
\]
Goal: Design ILP constraints $C$ and objective function $F$, s.t. maximizing $F$ subject to $C$ yields a “desirable” support graph

Variables define the space of “support graphs”
- Which nodes + edges between lexical units are active?

Objective Function: “better” support graphs = higher objective value
- Reward active units, high lexical match links, column header match, …
- Penalize spurious overuse of frequently occurring terms

Constraints
- ~50 high-level constraints
  - Basic Lookup, Parallel Evidence, Evidence Chaining, Semantic Relation Matching
  - Examples: connectedness, question coverage, appropriate table use
Evaluation

- 4th Grade NY Regents Science Exam
  - Focus on non-diagram multiple-choice (4-way)
  - 129 questions in completely unseen Test set
    - 6 years of exams; 95% C.I. = 9%
  - Score: 1 point per question (1/k for k-way tie including correct answer)

- Baselines:
  - **IR Solver**: Information Retrieval using Lucene search
    - Using 280 GB of plain text (50B tokens) “waterloo” corpus [AAAI, 2015]
    - IR Solver(Tables): Using same tables as TableILP
  - **PMI Solver**: Statistical correlation using pointwise mutual info.
    - Using 280 GB of plain text (50B tokens) “waterloo” corpus [AAAI, 2015]
  - **MLN**: Markov Logic Network, a structured prediction model
    - Using rules from 80K sentences [EMNLP, 2015]
Results: Same Knowledge

TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources.
Results

Ensemble performs 8-10% higher than IR baselines

Simple logistic regression. Features:  
- 4 from each solver’s score
- 11 from TableILP’s support graph (#rows, weakest edge, …)

(Clark et al, AAAI-2016)
Assessing Britteness: Question Perturbation

How robust are approaches to simple question perturbations that would typically make the question easier for a human?

- E.g., Replace incorrect answers with arbitrary co-occurring terms

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

<table>
<thead>
<tr>
<th>Solver</th>
<th>Original Score (%)</th>
<th>% Drop with Perturbation absolute</th>
<th>% Drop with Perturbation relative</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR</td>
<td>70.7</td>
<td>13.8</td>
<td>19.5</td>
</tr>
<tr>
<td>PMI</td>
<td>73.6</td>
<td>24.4</td>
<td>33.2</td>
</tr>
<tr>
<td>TableILP</td>
<td>85.0</td>
<td>10.5</td>
<td>12.3</td>
</tr>
</tbody>
</table>

More experiments in the paper!
How does it compare to SQL query?

- Database technology has more than 40 years history
  - How are you different from that?

Database world

Generate Tables

Query: Questions designed by humans, and for humans.

QA world

Generate Tables (based on training data)

Query: Look at your tables and generated compatible queries.

Time
Beyond tables: Tuple Inference

TableILP
Small number of semi-structured rows
Khashabi et al IJCAI’16

Tuple Inference
Large number of simple rows
Khot et al, ACL’17

- Inference over **independent rows**
- **Auto-generated short triples** will often lose critical context
- Additional structure present in the subject/object phrases
- Scaling to millions of tuples
- Matching on 100s of relations (e.g., ~150 in animal tensor)
Beyond tables: modelling compositionality

2,000 soldiers

Yet cold and starvation were not the most dangerous threats to soldiers at Valley Forge: Diseases like influenza, dysentery, typhoid and typhus killed two-thirds of the nearly 2,000 soldiers who died during the encampment. Dec 19, 2012

235 Years Ago, Washington's Troops Made Camp at Valley Forge ...

- One argument is a part of another.
  - The governor is a number
  - The object is a group modified by the governor.

Not all “of”s are the same
P: Teams are under pressure after PSG purchased Neymar this season. Chelsea purchased Morata. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: Who did Chelsea purchase this season?

A: {Alvaro Morata, Neymar, Romelu Lukaku}
Teams are under pressure after PSG purchased Neymar this season. Chelsea purchased Morata. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: *Who did Chelsea purchase this season?*

A: \{Alvaro Morata, Neymar, Romelu Lukaku\}

**Simple “lookup”** based on proximity to question words, answer type

- Basic word overlap suffices
- Neural methods (e.g., BiDAF) excel at
P: Teams are under pressure after PSG purchased Neymar this season. Morata, the recent acquisition by Chelsea, will start for the team tomorrow. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: Who did Chelsea purchase this season?

A: \{Alvaro Morata, Neymar, Romelu Lukaku\}

Simple rewording can confuse solvers
- E.g., BiDAF outputs “Neymar”
P: Teams are under pressure after PSG purchased Neymar this season. Morata, the recent acquisition by Chelsea, will start for the team tomorrow. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: Who did Chelsea purchase this season?

A: {Alvaro Morata, Neymar, Romelu Lukaku}

Linguistic understanding can help!
- Verbs, preposition, punctuation
- Domain agnostic => can use pre-trained NLP modules
Example, Rephrasing, simplified

Morata, the recent acquisition by Chelsea, will start for the team tomorrow.

- **Prepositional predicate: by (agent)**
  - The action done: the recent acquisition
  - Who/what did the action: Chelsea

- **Comma predicate: , (substitute)**
  - indicates an apposition structure

- **Verb predicate: purchase (agent)**
  - Purchaser (A0): Chelsea
  - Thing purchased (A1): Who

Identify the relation expressed by the predicate, and its arguments.
Linguistic understanding can help!

Create a **unified representation as a family of graphs**

- predicate-argument, trees, clusters, sequences

Instantiate with off the shelf NLP annotators

- Verb-SRL, Comma-SRL, Nom-SRL, Prep-SRL, Coref

A single representation is not enough to capture complexity of language
- **Augmented Graph** is the graph which contains potential alignments between elements of any two graphs.

- Connections via similarity / entailment
- Surface form well as semantics of their labels
SemanticILP: Example

Question: Who did Chelsea purchase this season?

Answer: Romelu Lukaku

Answer: Neymar

Answer: Alvaro Morata

Knowledge: ... Morata, the recent acquisition by Chelsea, will start for the team tomorrow...
SemanticILP

Translate QA into a search for an optimal subgraph

**Constraint:** Incorporate *global* and *local* constraints
  - **Global** e.g.
    - Have ends in question, paragraph and an answer
    - Connected graph
    - Exactly one answer
  - **Local** e.g.
    - If using a pred-arg graphs, use at least predicate and argument
    - If using a coref connected-comp. use at least two nodes

**Objective:** Capture what’s a valid reasoning, what’s preferred
  - **Preferences** e.g.
    - Use less sentences
    - Use sentences nearby
    - If using a pred-arg graph, give priority to the subject

Formulate as Integer Linear Program (**ILP**) optimization
  - Solution points to the best supported answer
Results #1: Aristo Questions

- Input: **Science question** Q with 4 answer options A
- Text: paragraph P obtained by concatenating top k Lucene-retrieved sentences for various answer options

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Regents 4th</th>
<th>Public 4th</th>
<th>Regents 8th</th>
<th>Public 8th</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exam scores</strong>, shown as a percentage</td>
<td>56.3</td>
<td>53.1</td>
<td>59.3</td>
<td>61.4</td>
</tr>
</tbody>
</table>
Results #2: ProcessBank [EMNLP-2014]

- Input: **Biology question** with 2 answer options

<table>
<thead>
<tr>
<th>Dataset</th>
<th>BiDAF</th>
<th>BiDAF tuned</th>
<th>IR</th>
<th>SyntProx Baseline*</th>
<th>ProRead* (structural supervision)</th>
<th>SemanticILP (linear comb. of components)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process Bank**</td>
<td>58.7</td>
<td>61.3</td>
<td>63.8</td>
<td>61.9</td>
<td>68.1</td>
<td>67.9</td>
</tr>
</tbody>
</table>

SemanticILP does not rely on domain-specific process structure annotation
- Close to the specialized, state-of-the-art ProRead system
- Substantially better than syntax-based and neural baselines

One single system tested on different datasets.

* Berant et al. (EMNLP, 2014)
** ~70% of the original dataset; true/false and temporal questions currently out of scope

More experiments in the paper!
Where are we?

How much “reasoning” do models do?

Neural networks
sophisticated word-matching
~ paraphrasing

PMI-solver
(Clark et al, 2016)

IR-solver (lucene)
(Clark et al, 2016)

TableILP
(Khashabi et al, 2016)

SemanticILP
(Khashabi et al, 2018)

TupleInf
(Khot et al, 2017)

MLNSolver
(Khot et al, 2015)

More Reasoning

How much do system decouple their language understanding from QA?

Neural networks
sophisticated word-matching
~ paraphrasing

TableILP
(Khashabi et al, 2016)

MLNSolver
(Khot et al, 2015)

SemanticILP
(Khashabi et al, 2018)

TupleInf
(Khot et al, 2017)

PMI-solver
(Clark et al, 2016)

language Understanding decompoled from QA

IR-solver (lucene)
(Clark et al, 2016)

More decoupled
For a “good” QA there is no notion of domain or dataset.

- Receive a question and give an answer

Many factors

- Reasoning shouldn’t be defined too narrowly
- Language understanding should be equated with training on datasets.

**Hard limit** for learn-with-training-only systems

- 10k training
- 100k training
- 1m training

We should push

More transferability

quantitative reasoning

reasoning
1. **Semi-structured inference**
   - We showed that can be very effective & robust
   - Goes beyond factoid-style QA
     - Chaining facts
     - First QA system to combine multiple semantic abstractions
     - State-of-the-art results on multiple datasets with different characteristics

2. **Reasoning** is a key to progress in QA.
   - No universal magic box
   - Decoupling QA and Language Understanding
   - Transferability is a key challenge
# CogCompNLP: most extensive NLP annotator

Many annotators: POS, NER, verb-SRL, comma-SRL, etc.

Code: [https://github.com/CogComp/cogcomp-nlp](https://github.com/CogComp/cogcomp-nlp)

<table>
<thead>
<tr>
<th>CogCompNLP (ours)</th>
<th>Sentence splitting</th>
<th>Tokenizing</th>
<th>Lemmatizing</th>
<th>Part of Speech tagging</th>
<th>Chunking</th>
<th>NER (4 labels)</th>
<th>Extended NER (18 labels)</th>
<th>Dependency Parsing</th>
<th>Quantity Parsing</th>
<th>Verb-sense Classification</th>
<th>Temporal Normalization</th>
<th>Mention Detection</th>
<th>Comma-SRL</th>
<th>Preposition SRL</th>
<th>Properbank (verb) SRL</th>
<th>Coreference resolution</th>
<th>SRL</th>
<th>Relation Extraction</th>
<th>Sentiment</th>
<th>OpenIE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Java</td>
<td>CORENLP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>OPENNLP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Python</td>
<td>SPACY</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>NLTK</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>TEXTBLOB</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
CogCompNLPy Light-weight Python NLP annotators

Many annotators: POS, NER, verb-SRL, comma-SRL, etc.
Code: https://github.com/CogComp/cogcomp-nlpy

```python
from sioux import remote_pipeline
pipeline = remote_pipeline.RemotePipeline()
doc = pipeline.doc("Hello, how are you. I am doing fine")

print(doc.get_lemma)
# will produce (hello Hello) (, ,) (how how) (be are) (you you) (. .) (i I) (be am) (do doing) (fine fine)

print(doc.get_pos)
# will produce (UH Hello) (, ,) (WRB how) (VBP are) (PRP you) (. .) (PRP I) (VBP am) (VBG doing) (JJ fine)
```
Motivation: Three Challenges

- **Diverse linguistic constructs** make QA systems brittle
  - Even the best systems are easily fooled by simple textual variations

- **Limited training data** in “interesting” QA domains
  - Paradigm of learning everything end-to-end doesn’t seem viable

- **Limited question understanding** in Aristo solvers
  - knowledge: explored several representations
  - question: still treated as tokens/chunks

**Goal:** Address these in the context of multiple-choice questions with supporting text, by reasoning over semantic abstractions of text
Majority of QA systems do “sophisticated” paraphrasing.
  - Not much reasoning

An experiment on a recent popular dataset SQuAD (Rajpurkar et al, 2016):

**Experiment**
- Take 50 questions and the paragraph contained its answer.
- Break the paragraph into sentences.
- Create (question, sentence) pairs, for any sentence in the paragraph
- For each (question, sentence) ask 3 people whether they can answer the question.
- Repeated it for all the sentences in each paragraph.

**Result**
- At least 2 (out of 3) people said they can answer 74% of questions, given a single sentence
- Manual inspection: The remaining questions all required co-reference reasoning.
John, a fast-rising politician, slept on the train to Chicago.

Verb Predicate: sleep
- Sleeper: John, a fast-rising politician
- Location: on the train to Chicago

Who was John?
- Relation: Apposition (comma)
  - John, a fast-rising politician

What was John’s destination?
- Relation: Destination (preposition)
  - train to Chicago

Identify the relation expressed by the predicate, and its arguments.
Extended Semantic Role Labeling

- Improved **sentence level** analysis; dealing with more phenomena.
  - Semantic role labelling
    - Events, Entailment, Winograd schemas
  - Abstract Meaning Representation (AMR) [Banarescu et al. 2013]
    - Expensive to produce large amounts of hand-annotated AMRs
    - Especially for other languages/genres
    - Limitations in terms of phenomena covered (hard to add more)
Teams are under pressure after PSG purchased Neymar this season. Morata is the recent acquisition by Chelsea. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: Who did Chelsea purchase this season?

A: {Alvaro Morata, Neymar, Romelu Lukaku}

Linguistic understanding can help!
- Verbs and their nominalization
- Domain agnostic => can use pre-trained NLP modules
Teams are under pressure after PSG purchased Neymar this season. Morata is the recent acquisition by Chelsea. The Spaniard looked like he was set for a move to Old Trafford for the majority of the summer only for Manchester United to sign Romelu Lukaku instead, paving the way for Morata to finally move to Chelsea for an initial £56m.

Q: Who did Chelsea purchase this season?

A: {Alvaro Morata, Neymar, Romelu Lukaku}

Simple rewording can confuse solvers

- E.g., BiDAF outputs “Neymar this season. Morata”
Matter takes up space and has mass. Two objects cannot occupy the same place at the same time.

Matter has properties (color, hardness, odor, sound, taste) that can be observed through the senses.

Objects have properties that can be observed, described, and measured: length, width, volume, size, shape, mass or weight, temperature, texture, reflectiveness of light.

Measurements can be made with standard units and nonstandard units.

The material(s) an object is made up of determines some specific properties of the object (sink/float, conductivity, magnetism).

Properties can be observed or measured with tools such as hand lenses, metric rulers, thermometers, balances, magnets, circuit testers, and graduated cylinders.

Objects and/or materials can be sorted or classified according to their properties.

Some properties of an object are dependent on the conditions of the present surroundings in which the object exists. For example: temperature - hot or cold; lighting - shadows, color; moisture - wet or dry.

Describe chemical and physical changes, including changes in states of matter.

Matter exists in three states: solid, liquid, gas.

Solids have a definite shape and volume.

Liquids do not have a definite shape but have a definite volume.

Gases do not hold their shape or volume.

Temperature can affect the state of matter of a substance.

Changes in the properties or materials of objects can be observed and described.
Relation Involving Which Objects?

Grouping of ~2500 key terms related to 4th grade science.
Semi-Structured Inference: Challenge #2

**Reasoning:** effective, controllable, scalable

**RULE solver [AKBC 2014]**
- forward chaining of logic rules
- Pros: easy to understand behavior (state space)
- Cons: focuses on *how* to search rather than *what* to look for

**MLN solver [EMNLP 2015]**
- approx. inference with probabilistic first-order logic
- Pros: “natural” fit, high-level specification
- Cons: inefficient, difficult to control, brittle with noisy input

**Integer Linear Programming (ILP) framework**
- *constraints and preferences*, *industrial-strength solvers*
Key components of the TableILP system contribute substantially to the eventual score.

<table>
<thead>
<tr>
<th>Solver</th>
<th>Test Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TableILP</td>
<td>61.5</td>
</tr>
<tr>
<td>No Multiple Row Inference</td>
<td>51.0</td>
</tr>
<tr>
<td>No Relation Matching</td>
<td>55.6</td>
</tr>
<tr>
<td>No Open IE Tables</td>
<td>52.3</td>
</tr>
<tr>
<td>No Lexical Entailment</td>
<td>50.5</td>
</tr>
</tbody>
</table>
Aristo’s Tablestore

- ~85 tables, ~10k rows, ~30k cells
- Defined with respect to questions, study guides, syllabus
ILP Complexity, Scalability

- ~50 high-level constraints

<table>
<thead>
<tr>
<th>Category</th>
<th>Quantity</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILP complexity</td>
<td>#variables</td>
<td>1043.8</td>
</tr>
<tr>
<td></td>
<td>#constraints</td>
<td>4417.8</td>
</tr>
<tr>
<td></td>
<td>#LP iterations</td>
<td>1348.9</td>
</tr>
<tr>
<td>Knowledge use</td>
<td>#rows</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td>#tables</td>
<td>1.3</td>
</tr>
<tr>
<td>Timing stats</td>
<td>model creation</td>
<td>1.9 sec</td>
</tr>
<tr>
<td></td>
<td>solving the ILP</td>
<td>2.1 sec</td>
</tr>
</tbody>
</table>

- **Speed:** 4 sec per question, reasoning over 140 rows across 7 tables
  - Contrast: **17 sec for MLN using only 1 rule** per answer option!
  - Commercial ILP engines (Gurobi, Cplex) much faster than SCIP
ILP Model

Operates on lexical units of alignment
- cells + headers of tables T
- question chunks Q
- answer options A

~50 high level constraints + preferences

**Variables** define the space of “support graphs” connecting Q, A, T
- Which nodes + edges between lexical units are active?

**Objective Function:** “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
ILP Model: Constraints

Dual goal: scalability, consider only meaningful support graphs

- **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)

- **Table Relevance Ranking**
  - TF-IDF scoring to identify top N relevant tables
Results: Exploiting Structured Knowledge

TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources.

Best of 3 MLN approaches:

A. First-order rules “as is”
   - Convenient, natural
   - Slow, despite a few tricks

B. Entity Resolution based MLN
   - Probabilistic “SameAs” predicate
   - Much faster, but brittle – low recall

C. Customized MLN: controlled search for valid reasoning chains
   - More controllable, more robust, more scalable (but still very limited)

[EMNLP-2015]
Two Approaches to Question Answering

- **Sophisticated physics model** of planetary movement
  - ✓ powerful model, would enable complex reasoning
  - ✗ difficult to implement, scale up, or learn automatically

- **Information retrieval / statistical association**
  - ✓ easy, generalizes well, often effective
  - ✗ limited to simple reasoning
  - ✗ expects answers explicitly written somewhere

Premise: a system that “understands” this phenomenon can correctly answer many variations!

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

New Zealand
WHY IS IT DIFFICULT?

One cannot simply map natural language to a representation that gives rise to reasoning.

Midas: I hope that everything I touch becomes gold.
**Ambiguity**

<table>
<thead>
<tr>
<th>Chicago was used by default for Mac menus through <strong>MacOS 7.6</strong>, and <strong>OS 8</strong> was released mid-1997.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chicago VIII</strong> was one of the early 70s-era <strong>Chicago</strong> albums to catch my ear, along with <strong>Chicago</strong>.</td>
</tr>
</tbody>
</table>

It's a version of **Chicago** – the standard classic **Macintosh** menu font, with that distinctive thick diagonal in the "N".

![Chicago Font Example](image1)

![Apple Logo](image2)

![MacOS Interface](image3)

![Chicago Band](image4)
Variability in Natural Language Expressions

Determine if Jim Carpenter works for the government

Jim Carpenter works for the U.S. Government.
The American government employed Jim Carpenter.
Jim Carpenter was fired by the US Government.
Jim Carpenter worked in a number of important positions.
 .... As a press liaison for the IRS, he made contacts in the white house.
Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter.
Former US Secretary of Defense Jim Carpenter spoke today...

Conventional programming techniques cannot deal with the variability of expressing meaning nor with the ambiguity of interpretation

Machine Learning is needed to support abstraction over the raw text, and deal with:
  - Identifying/Understanding Relations, Entities and Semantic Classes
  - Acquiring knowledge from external resources; representing knowledge
  - Identifying, disambiguating & tracking entities, events, etc.
  - Time, quantities, processes...
Where are we?

Knowledge decoupled from dataset

Reasoning
Aristo: Ensemble Approach

[AAAI-2016]