Reasoning-driven Question Answering

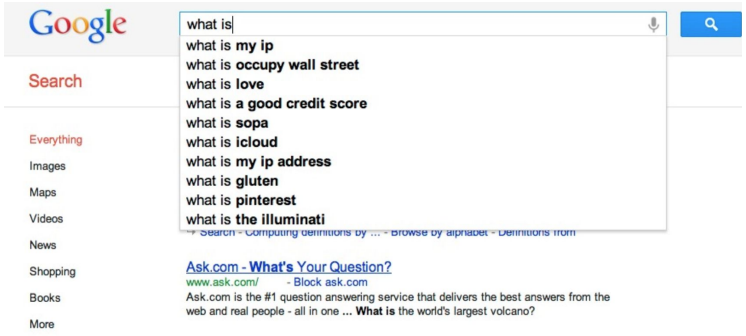
Daniel Khashabi

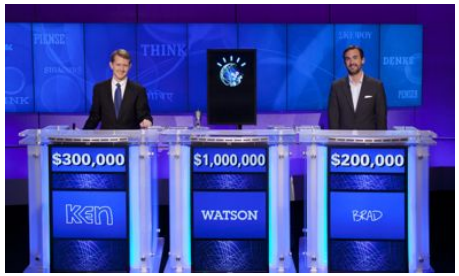
Committee:
Prof. Mitch Marcus (Chair), Prof. Zach Ives,
Prof. Chris Callison-Burch, Dr. Ashish Sabhrawal (external)

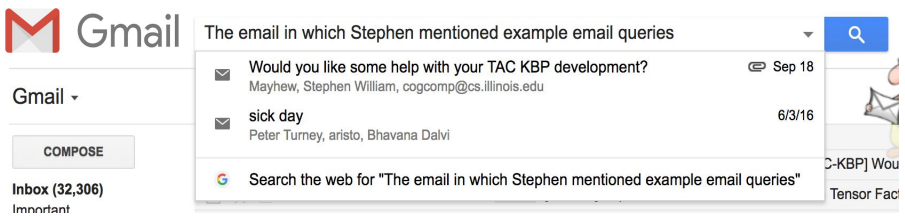
Thesis proposal meeting
Feb 9, 2018

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)

Programs with Commonsense

(John McCarthy, 1959)

Formalize world in **logical** form!

Example:
"My desk is at home" → at(I, desk)
"Desk is at home" → at(desk, home)

Hyp

Do **reasonin**

Exa
$\forall x \forall y \forall z$ at(x
∴ at(I, home)

McCarthy was right that, once you understand language you can do reasoning;
but he missed that NLU is difficult.

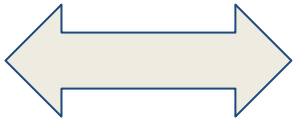
Hypothesis: Commonsense problems are solved by logical reasoning

What they missed: Variability and Ambiguity

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

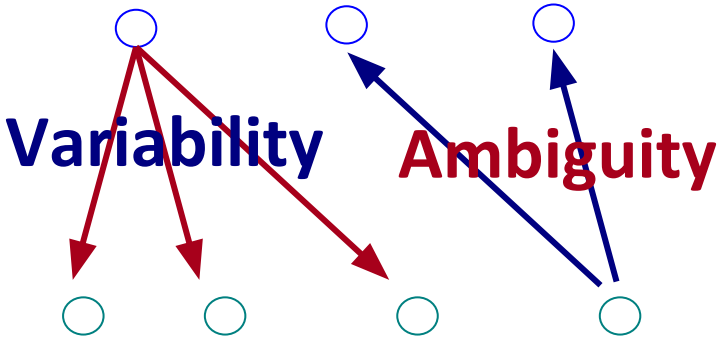
One cannot simply map natural language to a representation that gives rise to reasoning

"Chicago"



Meaning

Language

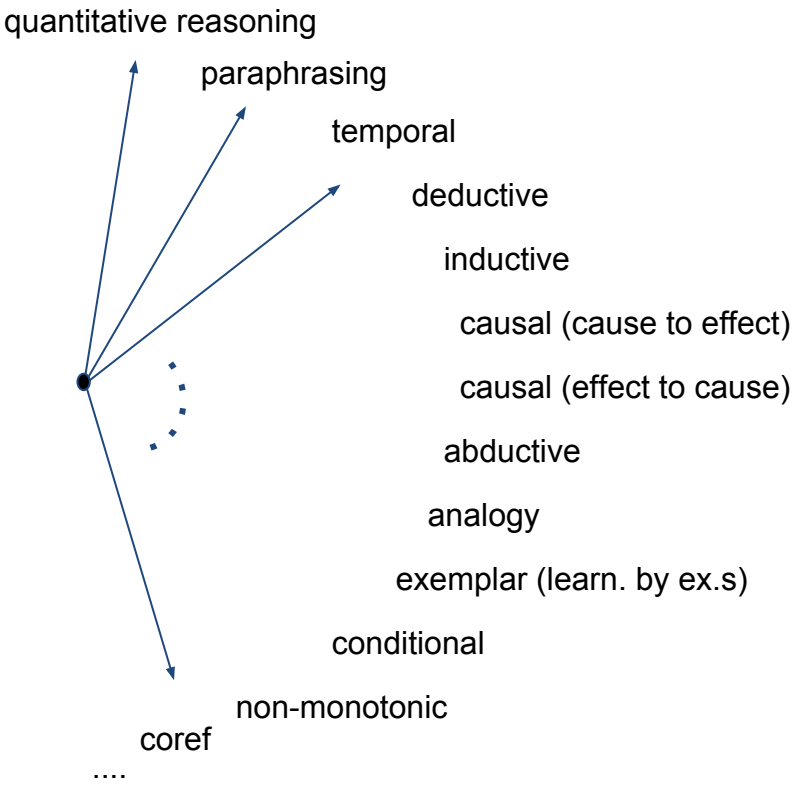


What they missed: Many faces of reasoning

- Reasoning is often studied in a very narrow sense.

Reasoning has many (infinite?) forms.

- Examples typically span multiple reasoning aspects.



Formal reasoning

- **Abductive reasoning**

Incomplete Observations → Best conclusion (maybe true)

(Bayesian Nets; Fuzzy Logic; Dampster-Shafer Theory)

The grass is wet, …
- It must have rained.
- Someone has watered them

- **Deductive reasoning**

General Rule → Specific conclusion (always true)

(modus ponens; modus tollens)

When it rains, objects get wet.
It rained.
- The grass must be wet.

- **Inductive reasoning**

Specific Observation → General Conclusion (maybe true)

Every time that it rains, the grass gets wet.
- It must be the case that with rain grass always get wet.

The many faces of reasoning

Q: When did Jack pass out?

The sunlight hit Jack and he passed out.
Options: morning, noon, night

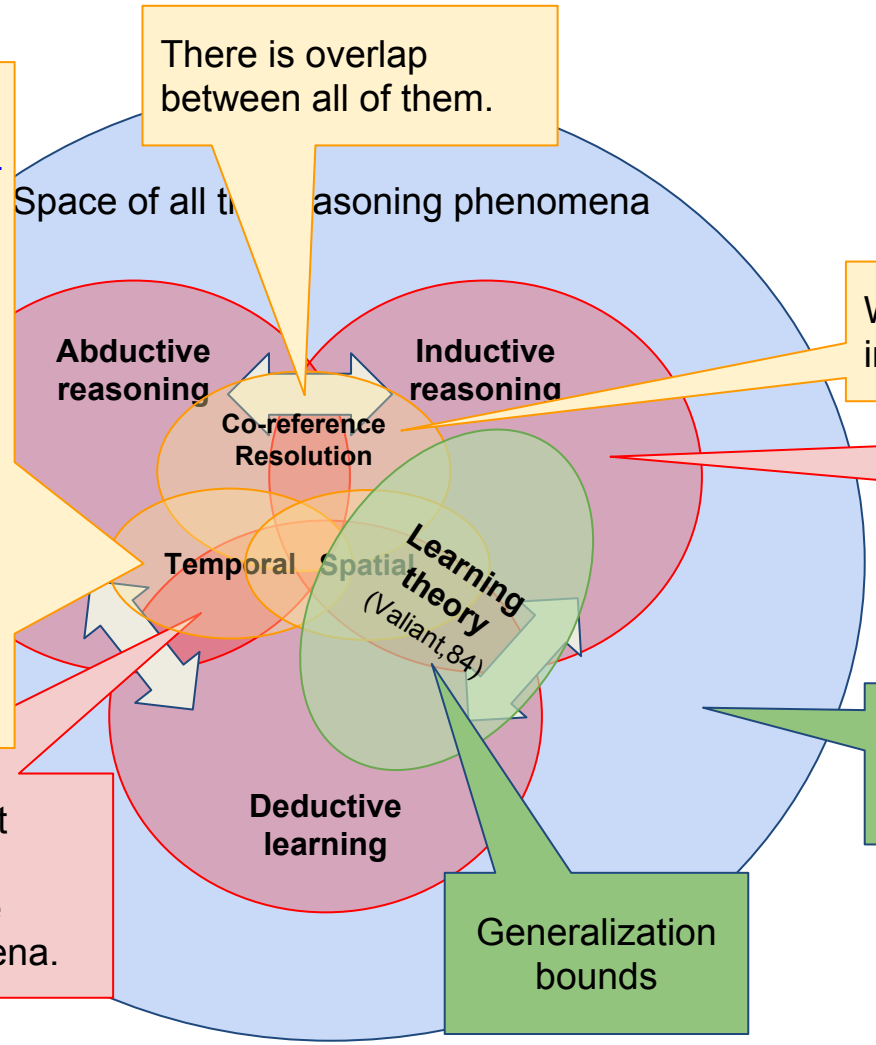
- "after" (temporal)
- "And" shows a temporal relation.
- "sunlight" can be:
  - morning; opening a window?

⇒ **Abduction: (probably) morning**

Jack passed out after the dinner.
Options: morning, noon, night

- "after" (temporal)
- "dinner" happens at night (temporal)
- how long is "dinner" (temporal)

⇒ **Deduction: night**



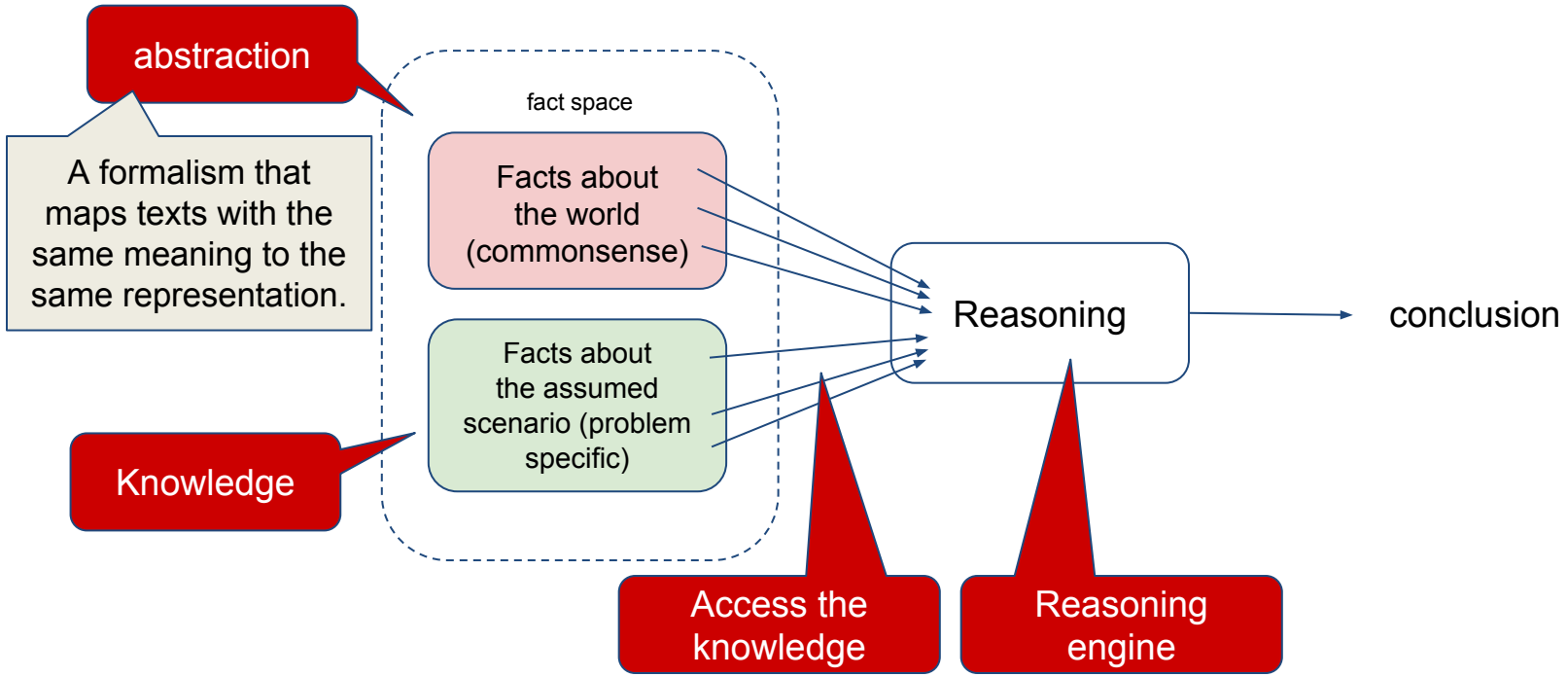
In *language*, things are not clearly disjoint.
⇒ An instance might have elements of both phenomena.

Thesis statement

Answering natural language questions, require a wide spectrum of reasoning abilities working together coherently, while affecting each other.
Focusing on creating this harmony and interplay is a key to making progress in natural language question answering.

The big picture

It's a convoluted subject and hard to define.



Roadmap

- Motivation
- **Background**
- **Previous work**
  - A formalism for abductive reasoning (IJCAI'16, AAAI'18)
  - Learning what to pay attention to in questions (CoNLL'17)
  - A dataset for reasoning over multiple sentences (submitted)
- **Proposed research**

Exams

Standardized science exams (Clark et al, 2015):

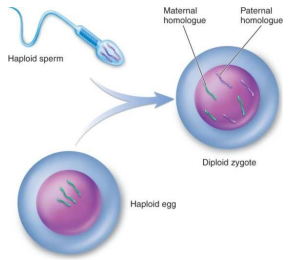
- Simple language; kids can solve them well, but they need to have the ability use the knowledge and abstract over it.

Q: Which physical structure would best help a bear to **survive a winter** in New York State?
A: (A) big ears (B) black nose (C**) thick fur** (D) brown eyes
P: … Polar bears, saved from the bitter cold by their thick fur coats**,** are among the animals in danger …

Biology exams (Berant et al, 2014):

- Technical terms and answer not easy to find.
- Requires understanding complex relations.

Q: What does meiosis directly produce?
(A) Gametes (B) **Haploid cells**
P: … Meiosis produces not gametes but haploid cells that then divide by mitosis and give rise to either unicellular descendants or a haploid multicellular adult organism. Subsequently, the haploid organism carries out further **mitoses, producing the cells** that develop into gametes.

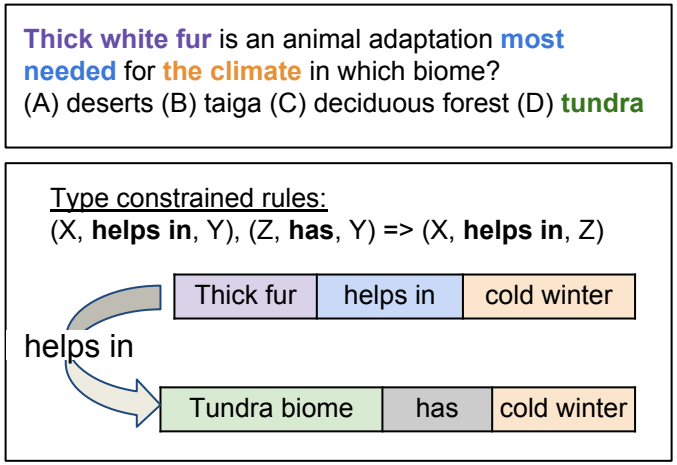


Penn
University of Pennsylvania

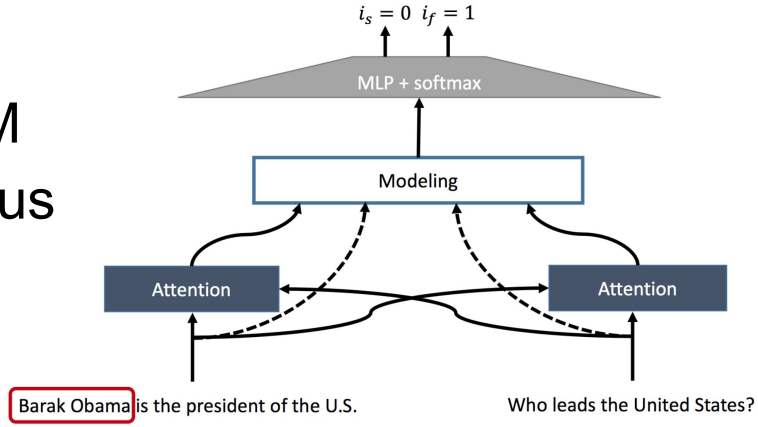
Evaluation: notable baselines

- IR **(Clark et al, AAAI'15)**
  - Information retrieval baseline (Lucene)
  - Using 280 GB of plain text

- TupleINF **(Khot et al, ACL'17)**
  - Inference over **independent rows**
  - **Auto-generated short triples**
  - And type-constrained rules



- BiDaF **(Seo et al, ICLR'16)**
  - Neural model: attention & LSTM
  - Extractive, i.e select a contiguous phrase in a given paragraph



Penn
University of Pennsylvania

Roadmap

- Motivation
- Background
- **Previous work**
  - A formalism for abductive reasoning (IJCAI'16, **AAAI'18**)
  - Learning what to pay attention to in questions (CoNLL'17)
  - A dataset for reasoning over multiple sentences (submitted)
- **Proposed research**

Semantic variability

Which physical structure would best **help a bear to survive a winter**?

(A) big ears (B) black nose (C) **thick fur** (D) brown eyes

Thick fur helps a bear survive a winter.

A thick coat of white fur helps bears survive in these cold latitudes.

Polar bears, saved from the bitter cold by their thick fur coats, are among the animals in danger of extinction because of the global warming and human activities.

A given “meaning” can be phrased many surface forms!

QA is a language understanding problem!

verb

Which physical structure would best help a bear to survive a winter?

(A) big ears (B) black nose (C) **thick fur** (D) brown eyes

comma

preposition

Polar bears, saved from the bitter cold by their thick fur coats, are among the animals in danger of extinction because of the global warming and human activities.

QA is fundamentally an NLU problem

A single abstraction is not enough

Collections of semantic graphs

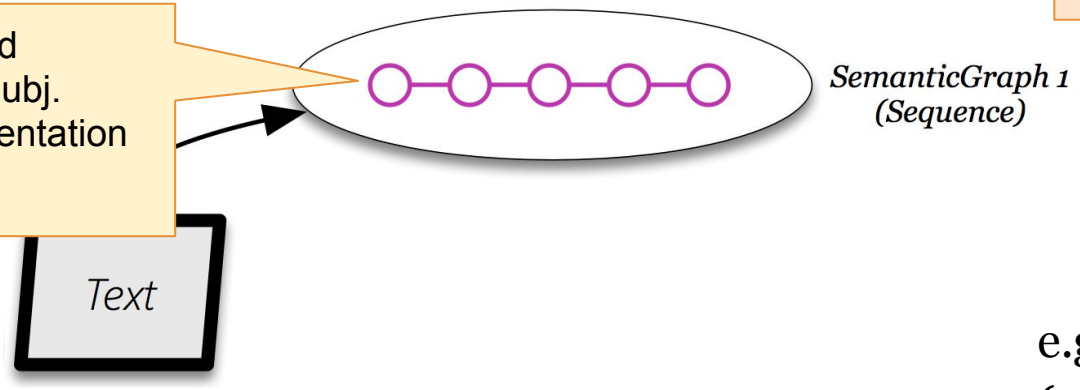
Create a unified representation of families of graphs

- predicate-argument, trees, clusters, sequences

A single representation is not enough to capture the complexity of language

- Surface word
- Label, e.g. subj.
- W2V representation

...



e.g named-entities

e.g dependency parse

e.g semantic role labeling (verb, preposition, comma)

e.g co-reference

e.g tables

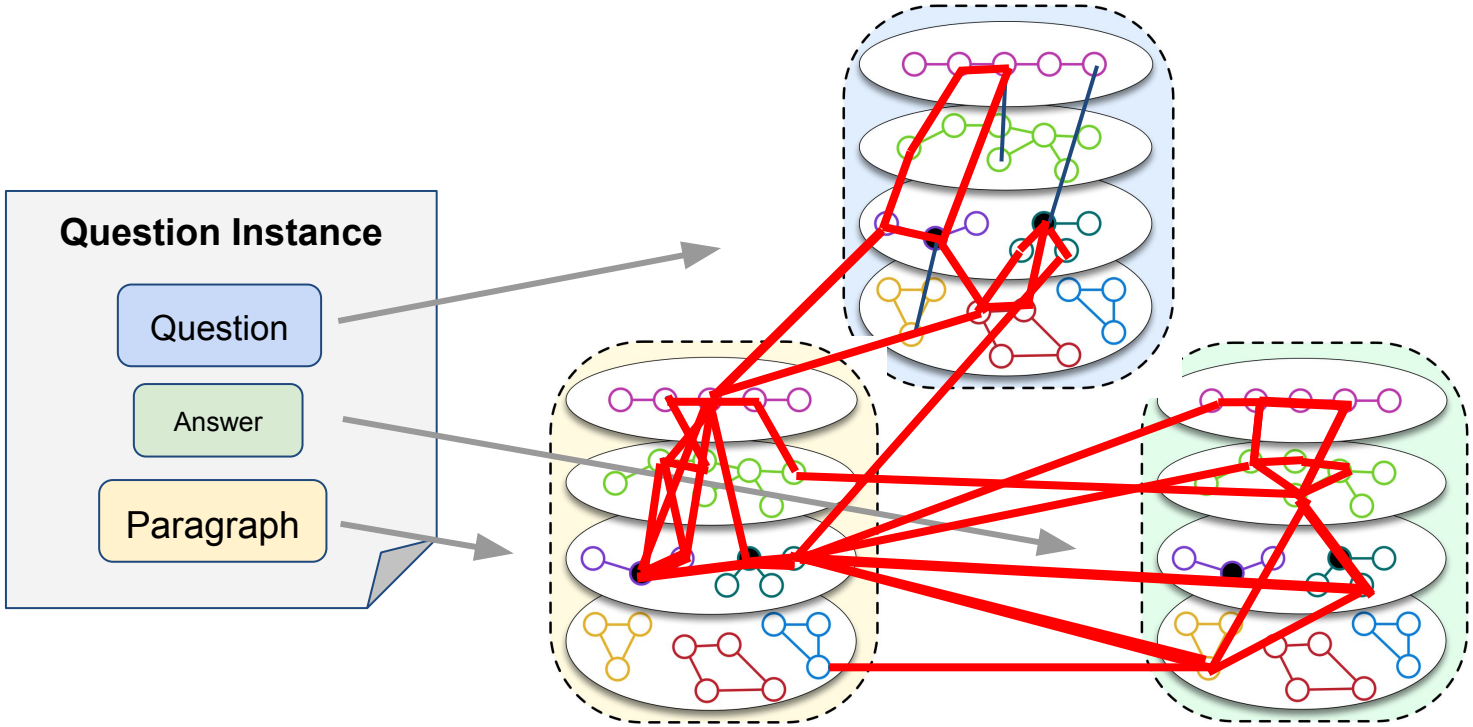
TableILP: IJCAI'16

Our representation has nothing to do with the QA task. It reflects our understanding of the language

Reasoning With a Meaning Representation

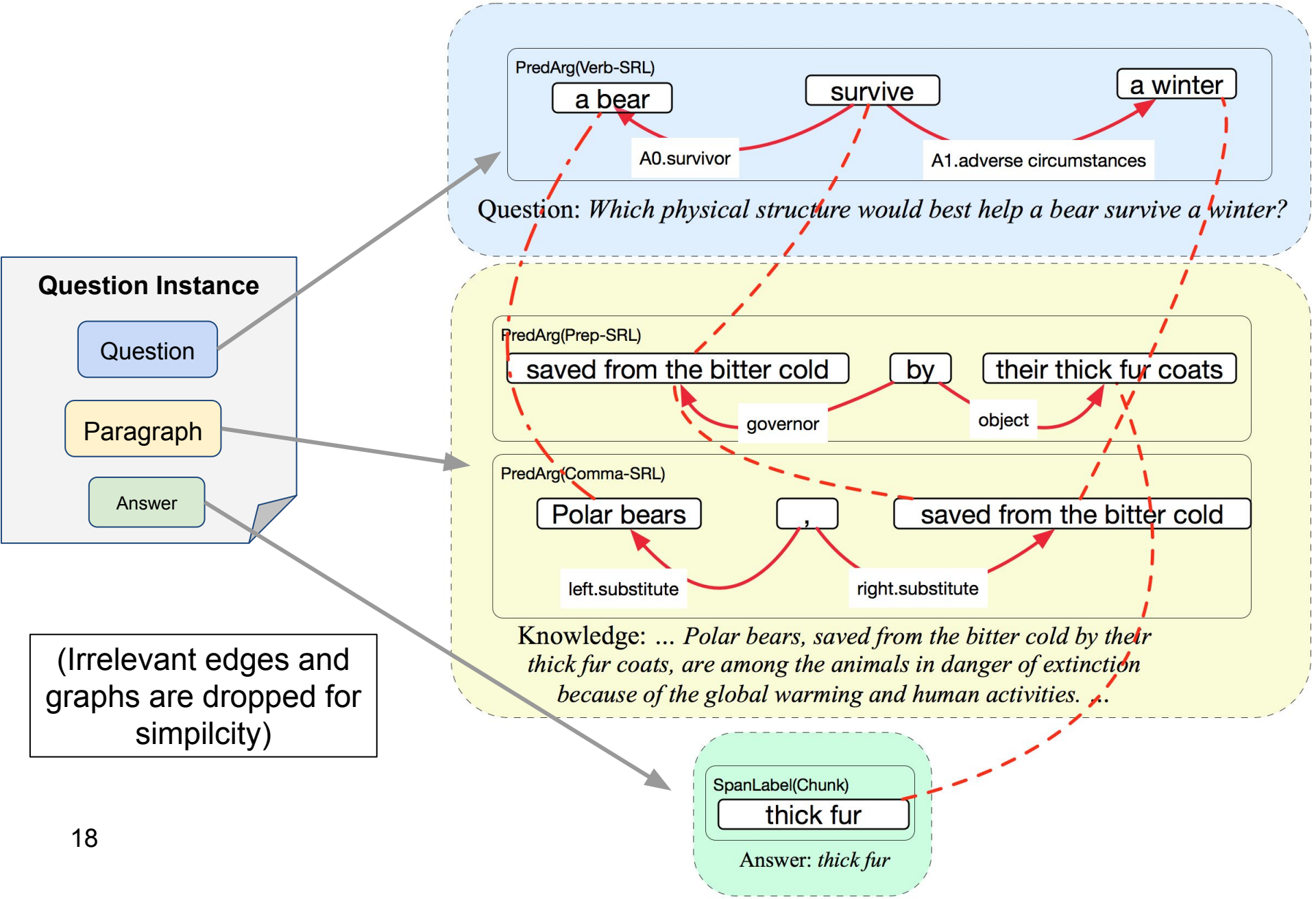
- **Augmented Graph** is the graph which contains potential alignments between elements of any two graphs

Connections via similarity / entailment



Reasoning formulated as best subgraph reasoning

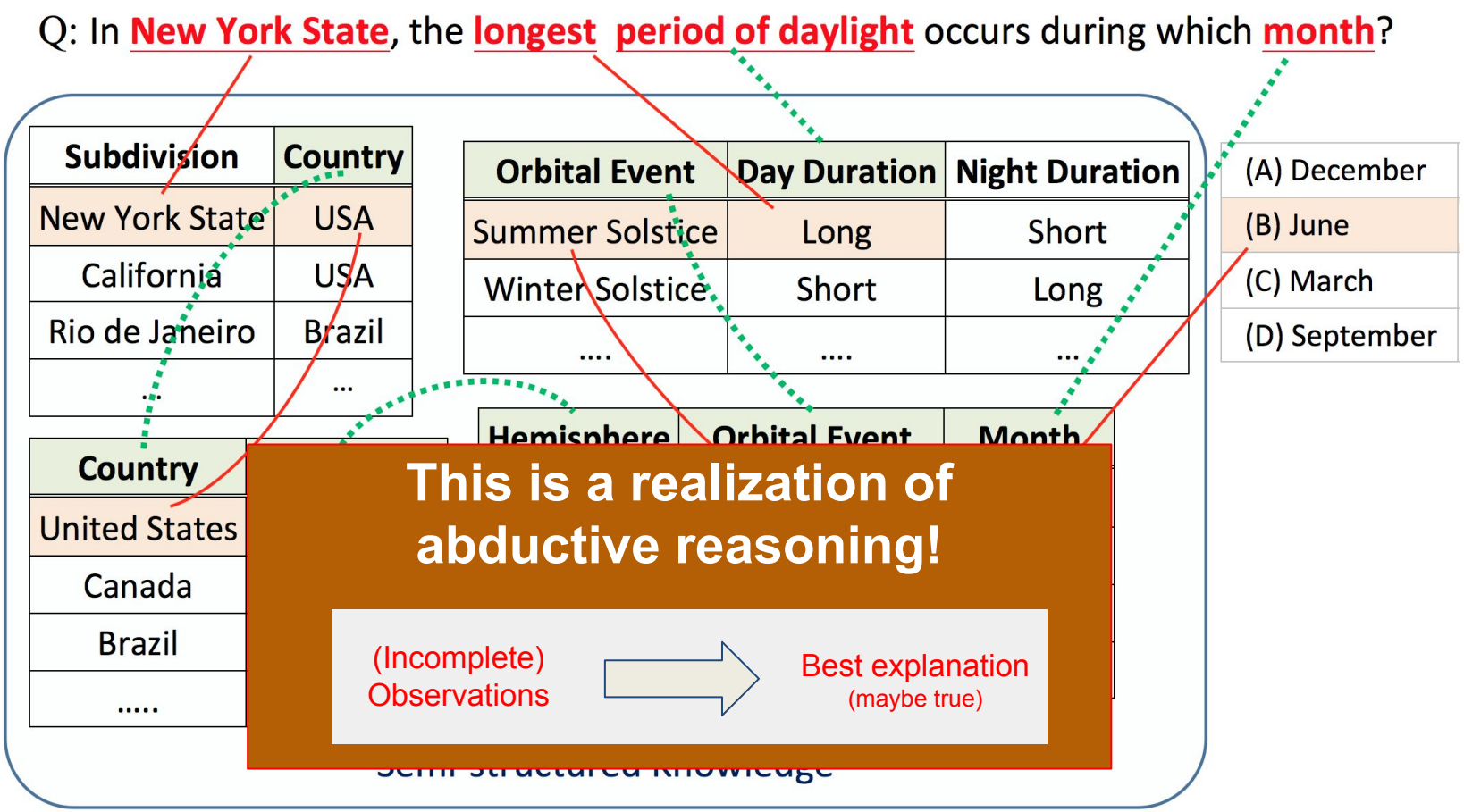
SamanticILP: Example subgraph



TableILP: Main Idea (IJCAI'16)

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

Link this information to identify the best supported answer!

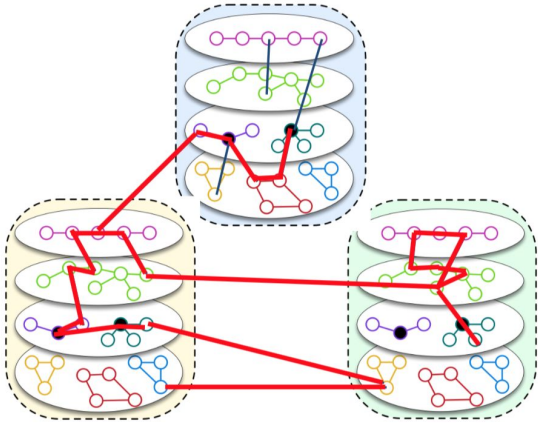


SemanticILP, some details.

Translate QA into a search for an optimal subgraph

Constraint: Incorporate **global** and **local** constraints

- **Global** e.g.
  - Have ends in question and paragraph
  - Connected graph
- **Local** e.g.
  - If using a pred-arg graphs,
    - use at least predicate and argument, or
    - use at least two arguments



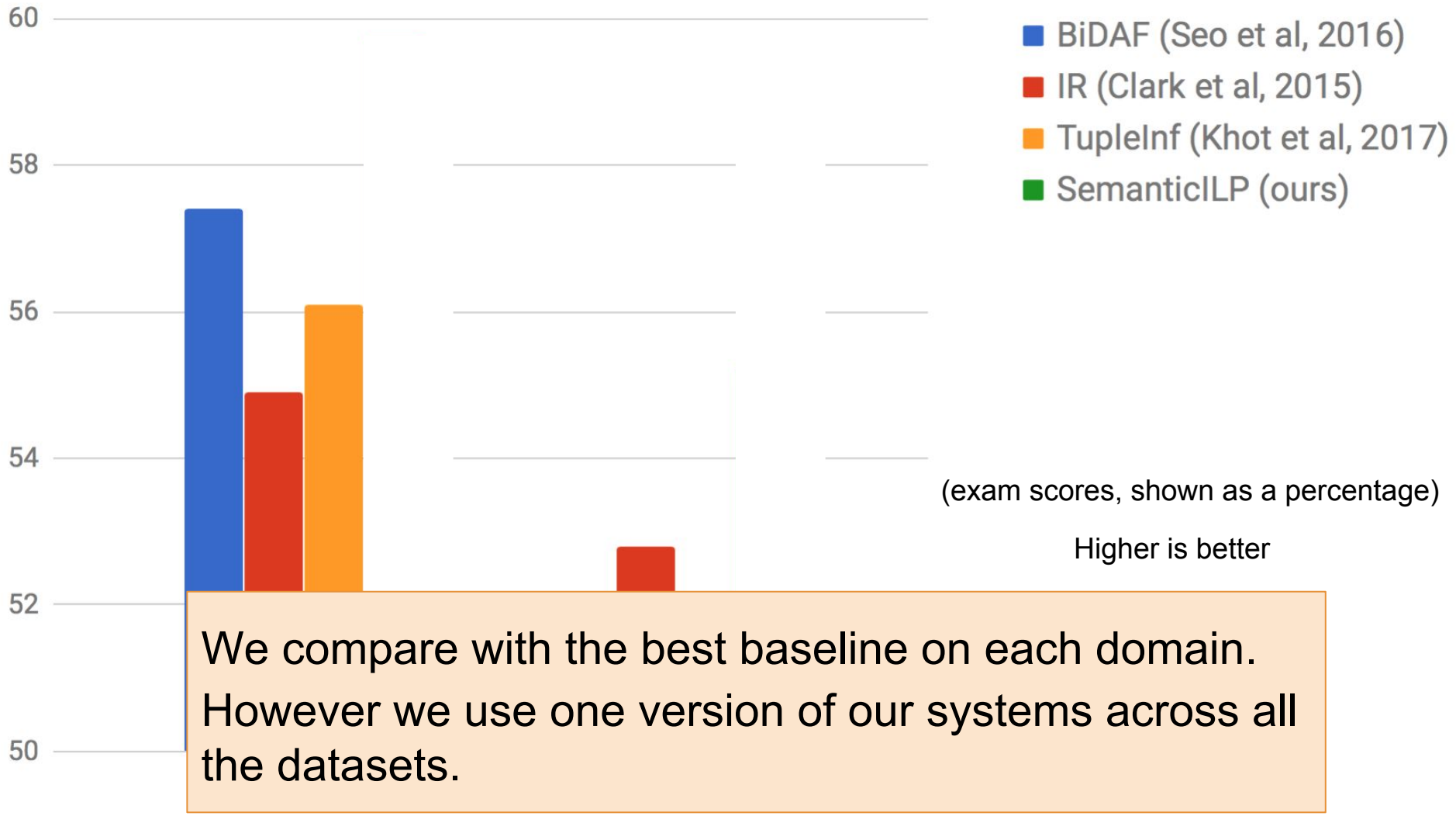
Objective: Capture what's preferred:

- **Preferences** e.g.
  - 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: Science Questions



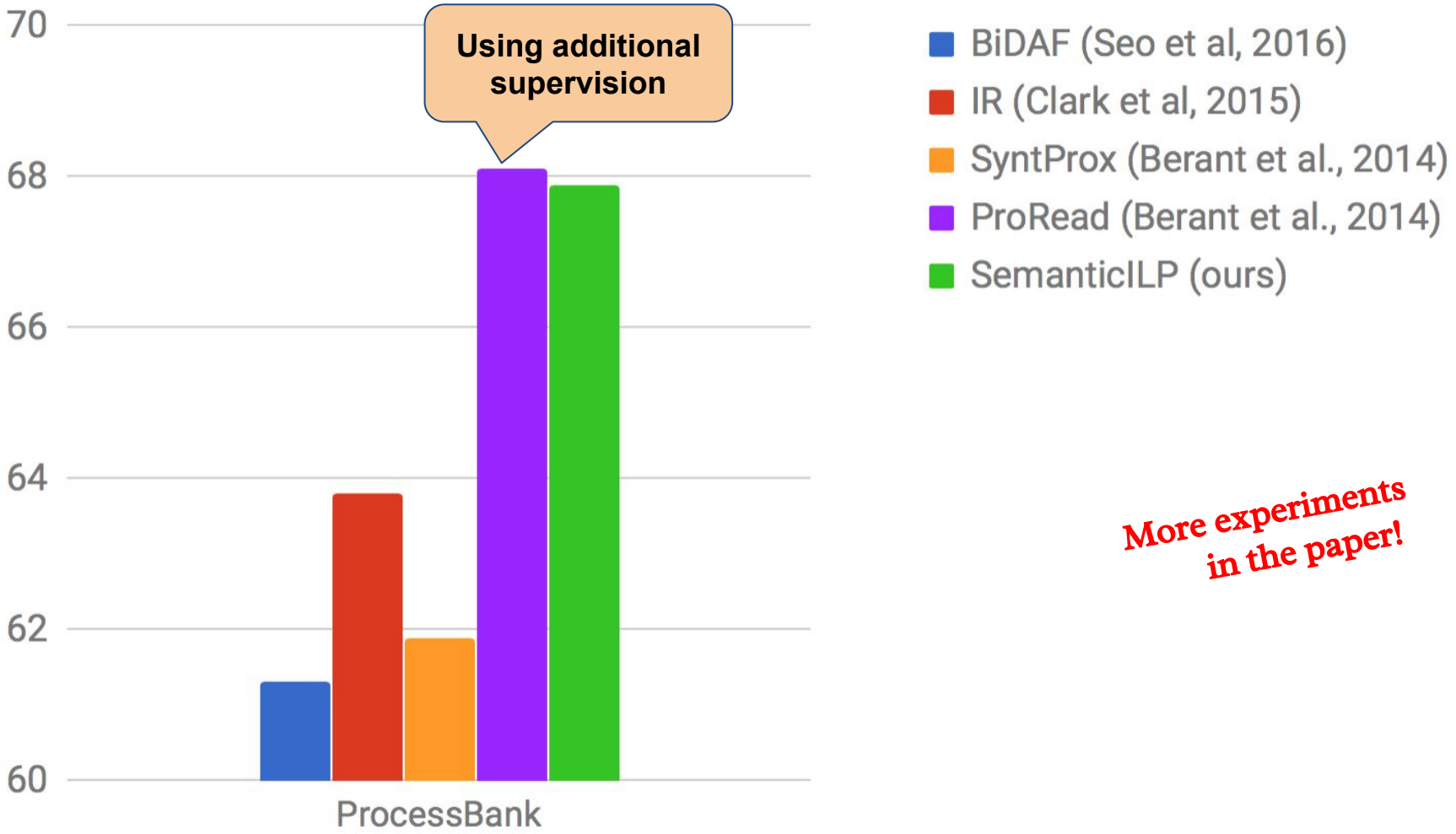
(exam scores, shown as a percentage)

Higher is better

We compare with the best baseline on each domain.

However we use one version of our systems across all the datasets.

Results #2: Biology Questions

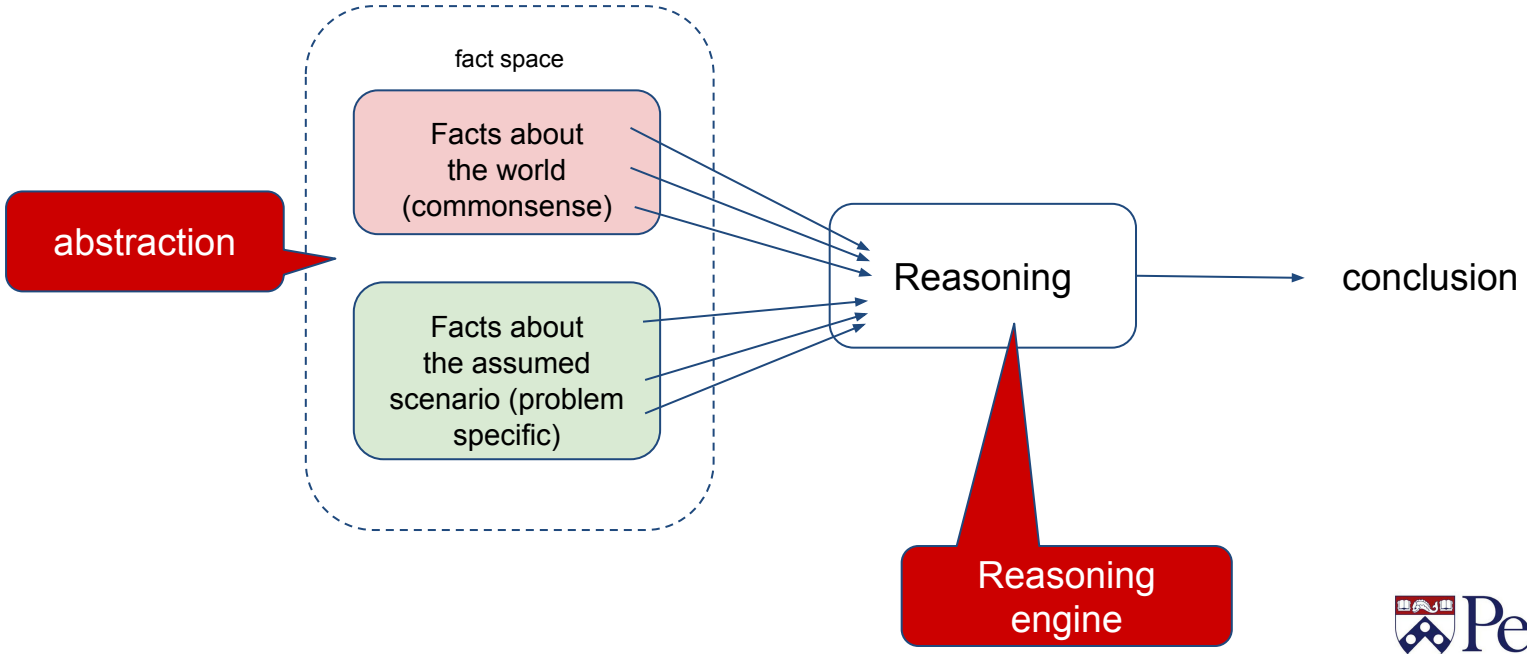


More experiments in the paper!

One single system tested on different datasets.

Summary

- Reasoning over language requires dealing with diverse set of semantic phenomena.
- Semantic variability ⇒ collection of semantic abstractions that are linguistically informed
- We decoupled "reasoning for QA" from "abstraction"
- Strong performance on two domains simultaneously



Roadmap

- Motivation
- Background
- **Previous work**
  - A formalism for abductive reasoning (IJCAI'16, AAAI'18)
  - Learning what to pay attention to in questions (CoNLL'17)
  - A dataset for reasoning over multiple sentences (submitted)
- **Proposed research**

Problem definition

As a part of QA reasoning engine, any system has to have an *attention mechanism* in reading questions (i.e. know what is important in questions)

animal
(A) find food (B) keep warmer (C) grow stronger (D) scape from predators

thicker
(A) find food (B) keep warmer (C) grow stronger (D) scape from predators

animal + thicker + hair
(A) find food (B) **keep warmer** (C) grow stronger (D) scape from predators

An animal grows thicker hair as a season changes. This adaptation helps to _____.
(A) find food (B) **keep warmer** (C) grow stronger (D) scape from predators

Original Question

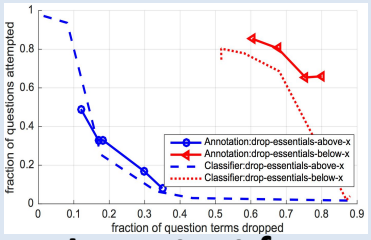
Challenge for QA systems: **Is a word in a question *important*, *redundant, or distracting?***

Overview of the approach

We introduce and study the notion of *essential question terms* with the goal of improving such QA solvers.

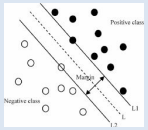
Essentiality in Questions

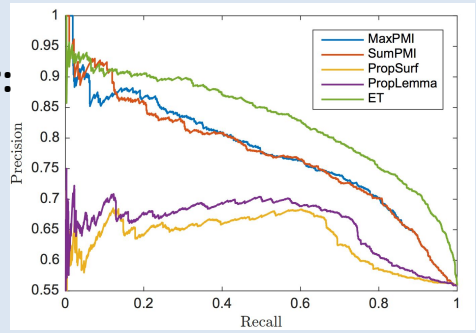
2K annotated questions
19K annotated terms



Important for humans!

State-of-the-art Essentiality classifier: F1 = 0.8, MAP = 0.9





Up to 5% increase in end-to-end QA performance

Crowd-Sourced Essentiality Dataset

- Collected ~2k science exam questions for the annotation.
- Questions annotated by 5 workers, resulted in ~20k annotated terms.

Instructions

Below is an elementary science question along with a few answer options. Using checkboxes, tell us which words or phrases of the question are essential for choosing the correct answer option, keeping in mind that:

- Essential phrase will change the core meaning.
- Non-essential item will not change the answer.
- Grammatical correctness is not important.

Examples

1. Which type of energy does a person use to pedal a bicycle? (A) light (B) sound (C) mechanical (D) electrical
2. A turtle eating worms is an example of (A) breathing (B) reproducing (C) eliminating waste (D) taking in nutrients
3. A duck's feathers are covered with a natural oil that keeps the duck dry. This is a special feature ducks have that helps them (A) feed their young (B) adapt to the environment (C) attract a mate (D) search for food

Mark the essential words:

How ☐ does ☐ the ☐ length ☐ of ☐ daylight ☐ in ☐ New ☐ York ☐ State ☐ change ☐ from ☐ summer ☐ to ☐ fall ☐ 1) It decreases. 2) It increases. 3) It remains the same.

Validity of the collected scores

An extra step to validate our hypothesis.

Specifically, we create a challenge for annotators by dropping terms, that:

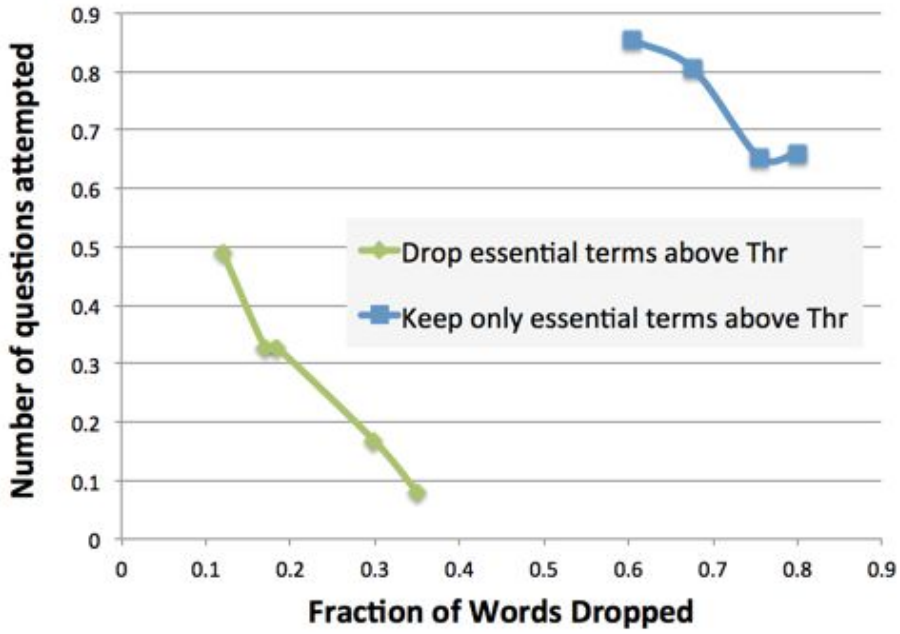
- **Have the high essentiality scores.**
- **Have the low essentiality scores.**

And ask them **whether they can answer the question or not**.

An **** grows thicker hair as a season changes. This **** helps to _____.
(A) find food (B) keep warmer (C) grow stronger
(D) scape from predators (E) I don't know. The information is not enough

Hypothesis:

- with dropping essential terms, humans would answer **very little** questions.
- with dropping non-essential terms humans still can answer **majority** of questions.



An *Essential-Term* Classifier

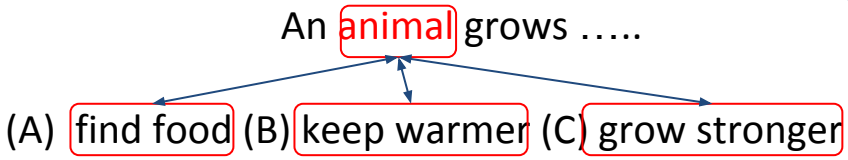
- Trained a linear SVM classifier
  - Real-valued essentiality scores are binarized
  - Features include
    - Syntactic (e.g., dependency parse based)
    - Semantic (e.g., Brown cluster representation of words)
    - As well as their combinations.
    - In total, we use 120 types of features.

Is this problem even *learnable*?

- **Supervised baselines:**
  - **PropSurf and PropLem:** Score for a term is proportional to times it was marked as essential in the annotated dataset.

Hypothesis: Essentiality is a function of context.

- **Unsupervised baselines:**
  - **MaxPMI and SumPMI:** score the importance of a word x by max-ing or summing, resp., PMI scores *p(x, y)* across all answer options y for q.

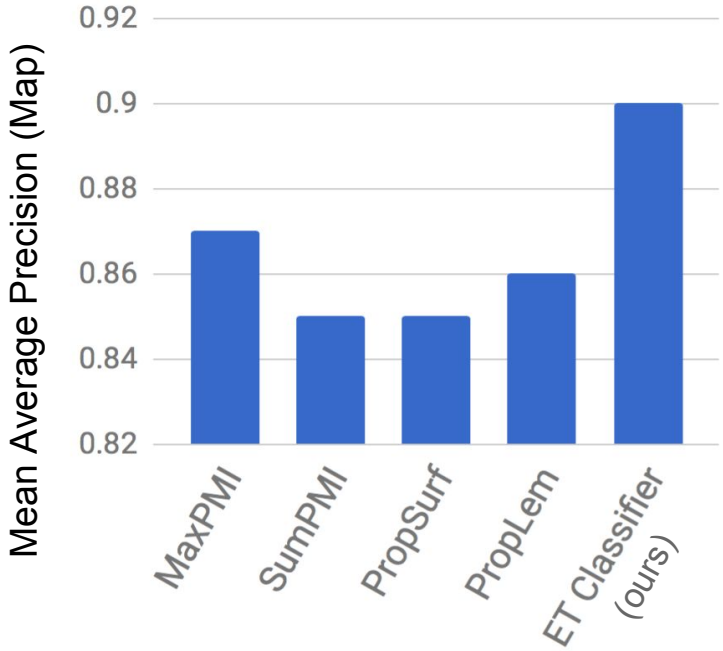


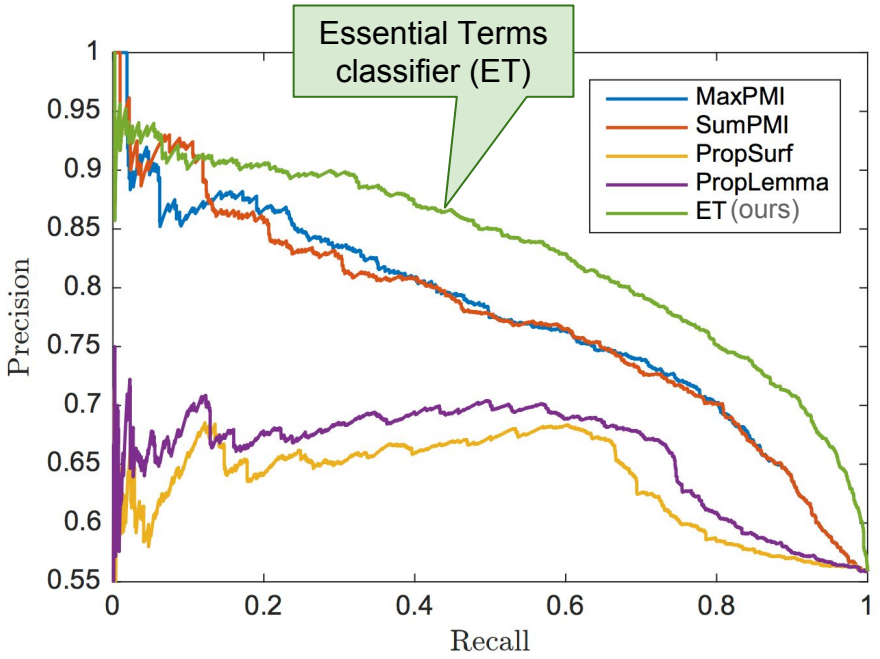
A proxy for how relevant two terms are, based on lots of unsupervised data.

An *Essential-Term* Classifier, contd.

Binary Classification of Terms.

- Our classifier performs significantly better than the baselines.





Ranking Question Terms.

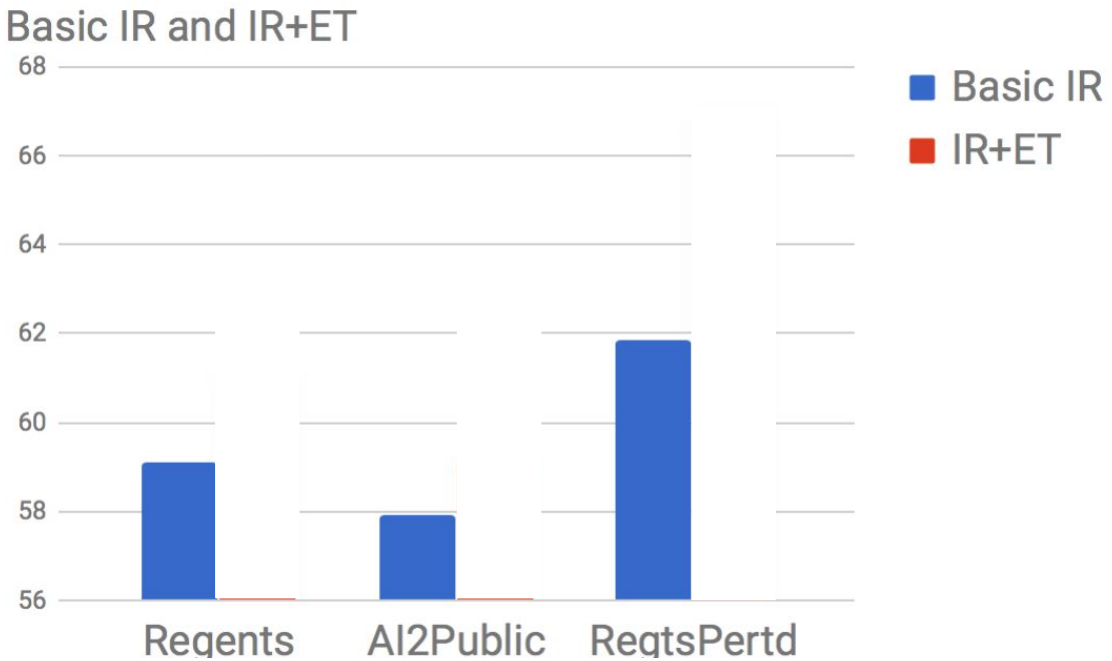
- Rank all terms within a question in the order of [essentiality] score.

Penn
University of Pennsylvania

End-to-end systems + Essential Terms

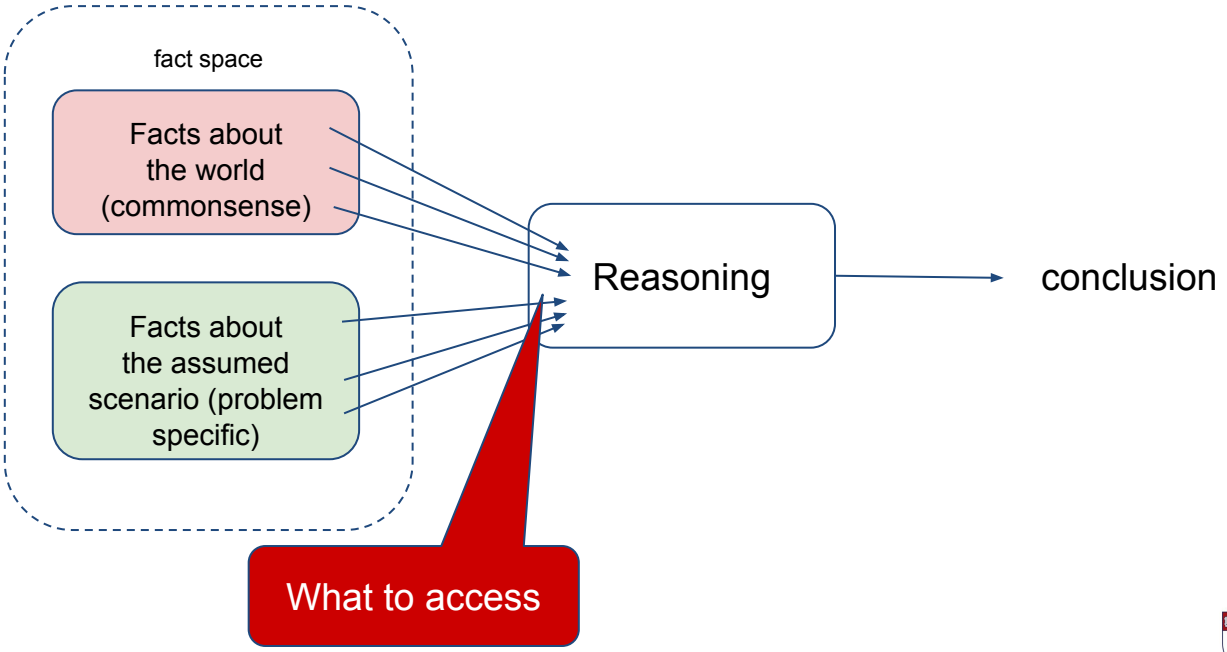
IR + Essential Terms Classifier

Instead of querying (q, a) pair, we query (q', a), with q' being subset of q, which has essentiality score above some threshold.



Summary

- There is a need to understand *what is important* questions.
- We introduce and study the notion of *essential question* terms with the goal of improving such QA solvers.



Roadmap

- Motivation
- Background
- **Previous work**
  - A formalism for abductive reasoning (IJCAI'16, AAAI'18)
  - Learning what to pay attention to in questions (CoNLL'17)
  - A dataset for reasoning over multiple sentences (submitted)
- **Proposed research**

MultiRC: Reasoning over multiple sentences.

A reading comprehension challenge set with questions that require 'reasoning' over more than one sentences in order to answer

S1: Most young mammals, including humans, play.
S2: Play is how they learn the skills that they will need as adults.
S6: Big cats also play.
S8: At the same time, they also practice their hunting skills.
S11: Human children learn by playing as well.
S12: For example, playing games and sports can help them learn to follow rules.
S13: They also learn to work together

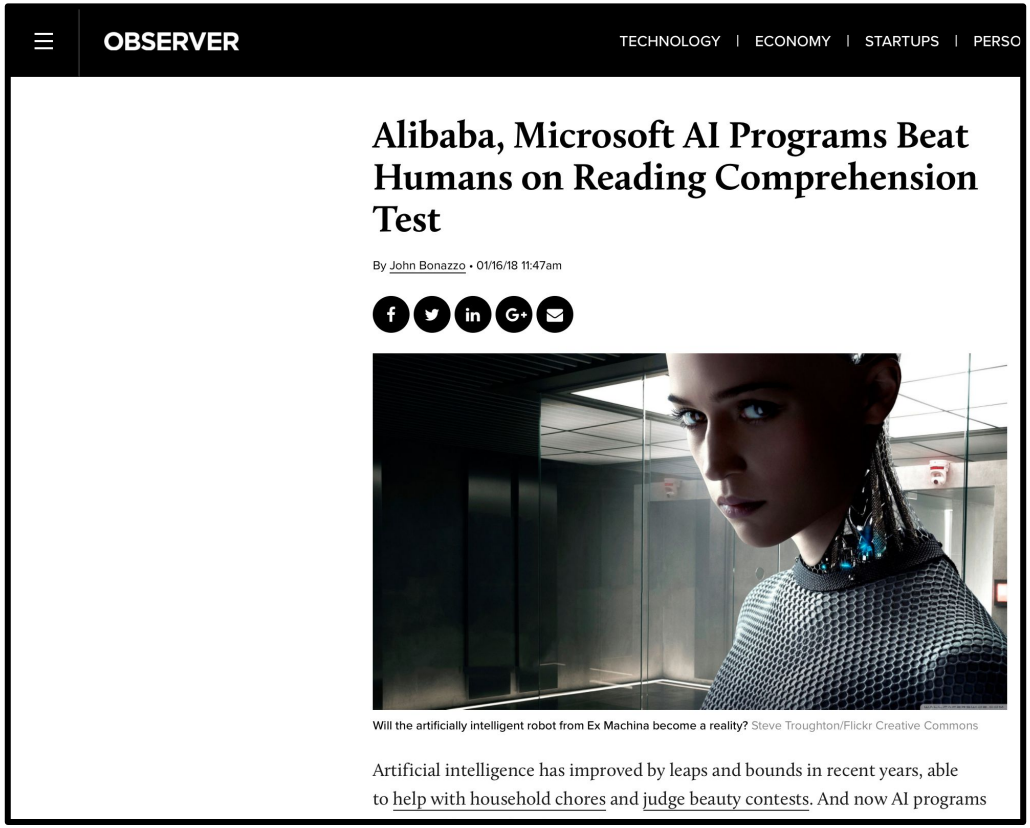
What do human children learn by playing games and sports?
A)* They learn to follow rules and work together
B) hunting skills
C)* skills that they will need as adult

Requires multiple sentences.

Number of correct answers not specified
(finding correct answers vs finding the most-correlated response)

Why do we need yet another RC dataset?

- **Datasets are often easy to solve.**
  - Most datasets are relatively easy and can be ‘solved’ with simple lexical matching.
  - >75% of SQUAD questions can be answered by the sentence that is lexically most similar to the question



Why “multi-sentence” questions?

There are efforts to design “reasoning-forcing” challenges

A prominent example:

- bAbI (Weston et al, 2015): small dataset on 10 tasks (reasoning forms).
- Issue: reasoning-specific questions (templated text).

Too much restriction

While not making too restricted assumptions, we want to define a proxy for reasoning content of questions.

“Multi-sentence” assumption:

- Does not restrict us to a narrow class of “reasoning” phenomena
- While forcing questions to have something more than trivial

Verifying multi-sentence-ness of questions

Given a sentence and a question, answer if the question can be answered

If turkers say “yes”, for at least one sentence → the question is not multi-sentence

Instructions

Answering Questions

You will be shown a sentence and a question. For each question,

- You have to say whether (Yes/No) the information provided in the sentence is enough to answer the question. If the answer is yes, you have to write the correct answer.
- When saying Yes/No **do not use any background knowledge.** Only use the information given in the sentence.

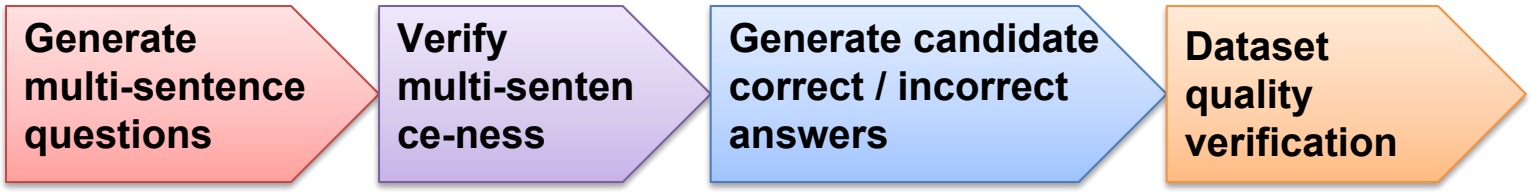
Below are a few example sentences and questions (and answers).

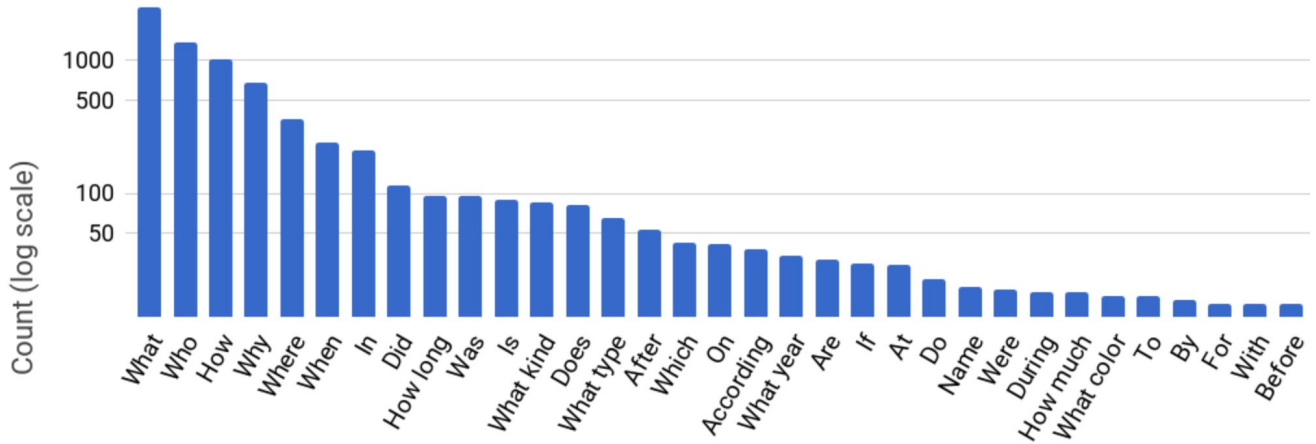
Sentence: GOP leaders submitted the new offer Tuesday afternoon in an effort to appease Democrats, whose votes are needed to avert a shutdown of federal agencies, several House and Senate aides said.
Question: Who has to be appeased to keep the government open?
Can the above question be answered using only the information provided in the given sentence?
(•) Yes; the information provided in the sentence is enough to answer the question.
( ) No; the information provided in the sentence is **not** enough to answer the question.
Answer:
the Democrats

Explanation: The sentence says that "the Democrats" have to be appeased, which answers the question.

MultiRC: Question generation pipeline

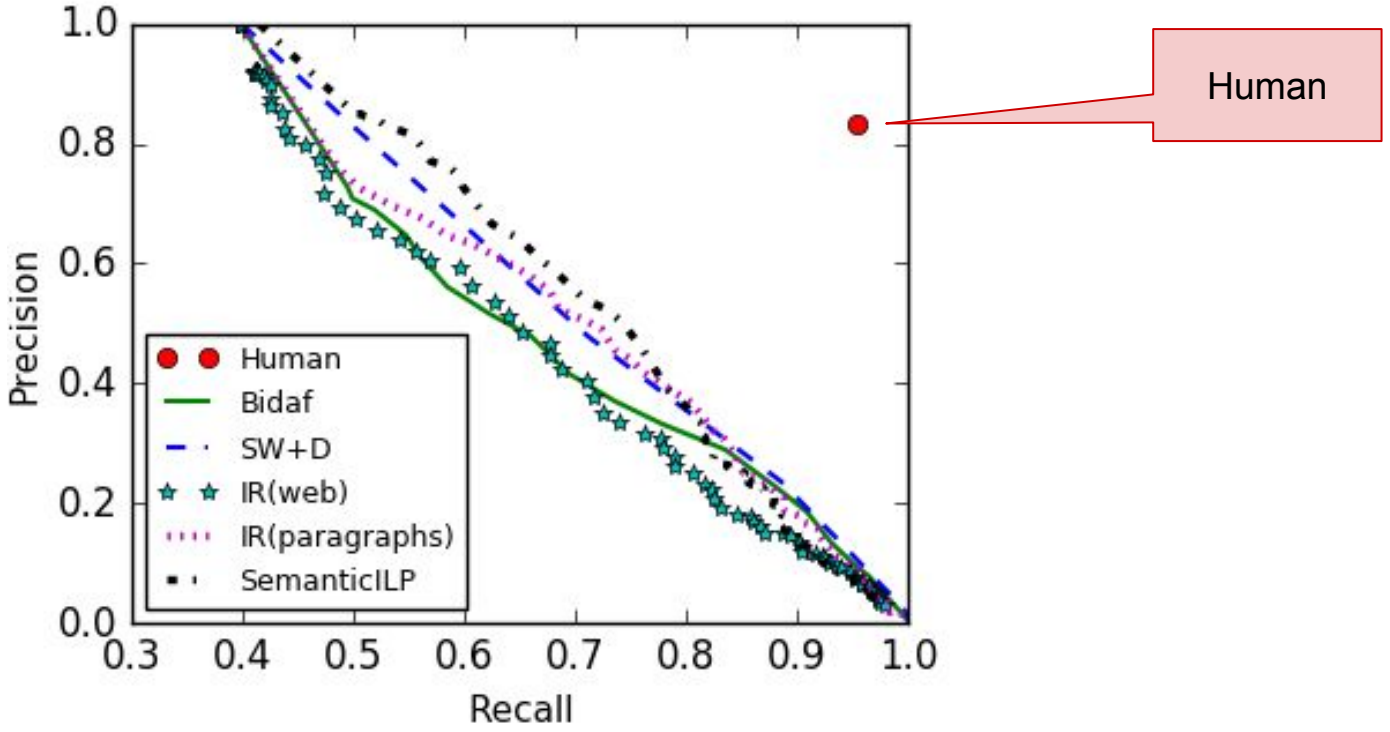
- +10,000 questions (6.5k are multi-sentence)
- on +700 paragraphs
- From 8 domains (fictions, news, science, social articles, Wikipedia, ...)





Baseline performances

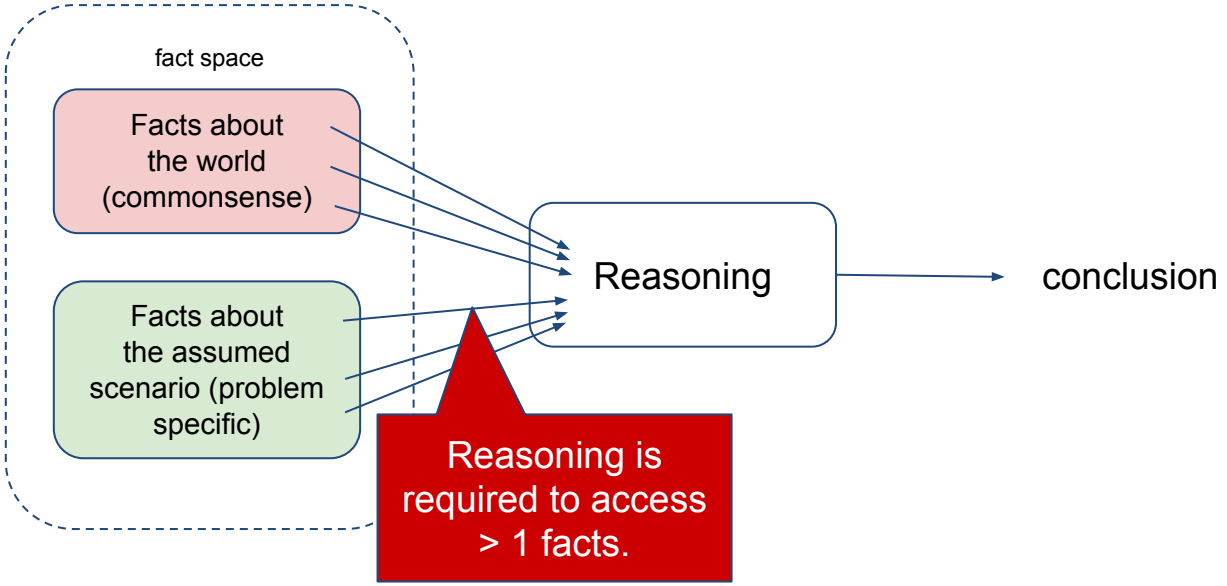
- Predict real-valued score per answer-option.
- For a fixed threshold, select answer-options that have score above it.



Summary

- We need reading comprehension playground which requires deeper “reasoning”

- An approach proposed here: enforcing dependence on multiple sentences.

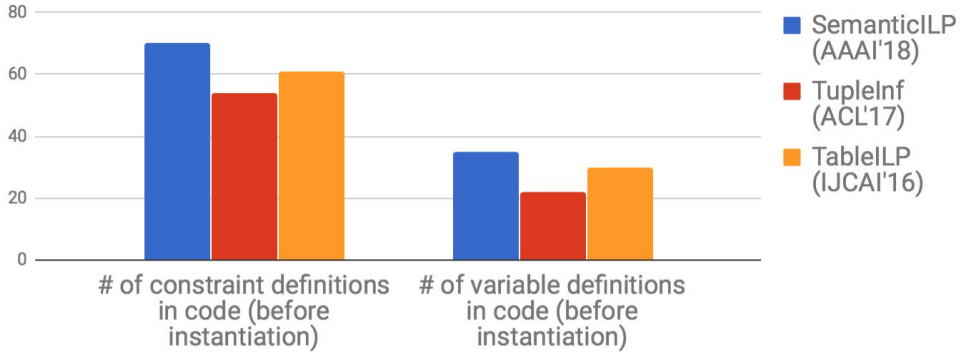


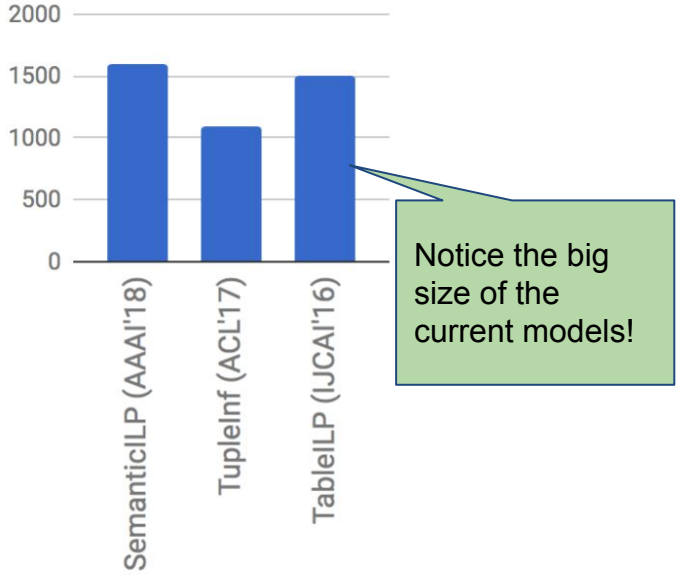
Roadmap

- Motivation
- Background
- Previous work
  - A formalism for abductive reasoning (IJCAI'16, AAAI'18)
  - Learning what to pay attention to in questions (CoNLL'17)
  - A dataset for reasoning over multiple sentences (submitted)
- **Proposed research**

Proposal 1: a programming language for reasoning over abstractions of NL

Recent models are highly complex in design





- Hard to modify and extend
- No one is going to continue developing it

Despite the apparent complexity:

- There is much redundancy
- Some behaviors could be described with a higher level abstraction

The models are actually describable in less than a page, when described in English!

Proposal 1: a programming language for reasoning over abstractions of NL

The current systems are inaccessible to other researchers.

Create a high-level programming language to model reasoning over semantic abstractions of natural language.

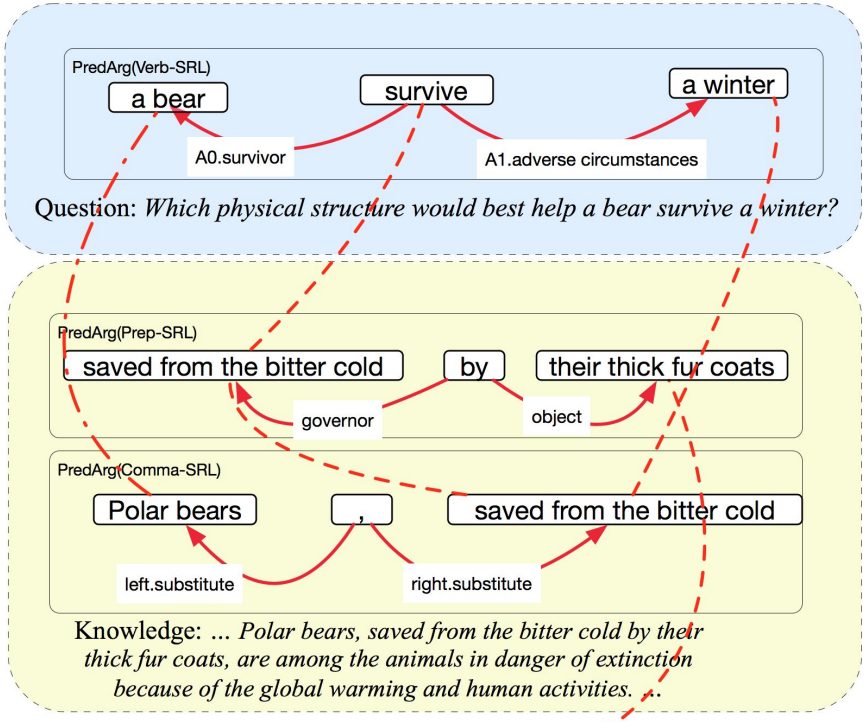
Expected features:

- ➔ Bring in knowledge in user's data-structures, or use the internal ones
- ➔ Define general properties of the "reasoning" alignment
- ➔ Define scoring mechanism for edge alignments
- ➔ Higher-level definitions for addition constraints:
  - Inside each connected component
  - Between the connected components
  - The whole alignment

Proposal 1: a programming language for reasoning over abstractions of NL

Example: for any (verb/nominal) predicate-argument graph, use the predicate and at least an argument (if not, don't use it).

ILP-level implementation of this requires at least 30 lines of code!




```
predicateArgumentGraph.
     addConstraint(atleast 1 predicate and atleast 1 argument)
```

Evaluation

- Reimplement previous systems, in a much simpler language.
- Show easy extensions; e.g. unifying TableILP + TupleILP

Proposal 2: QA "robustness" and "generalization"

Commonly accepted that QA systems are often brittle, because they fail with small *variations*.

Replace incorrect answers with arbitrary co-occurring terms

IJCAI'16

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	relative
IR	70.7	13.8	19.5
PMI	73.6	24.4	33.2
TableILP	85.0	**10.5**	**12.3**

Recent studies show that the models trained on these datasets don't do much of 'reasoning' (Jia & Liang, EMNLP'17)

Proposal 2: QA "robustness" and "generalization"

The brittleness in the current systems is due to not using the "right" paradigm.

dual

A QA system that answers a question for the "right" reason, would not fail with small *variations*.

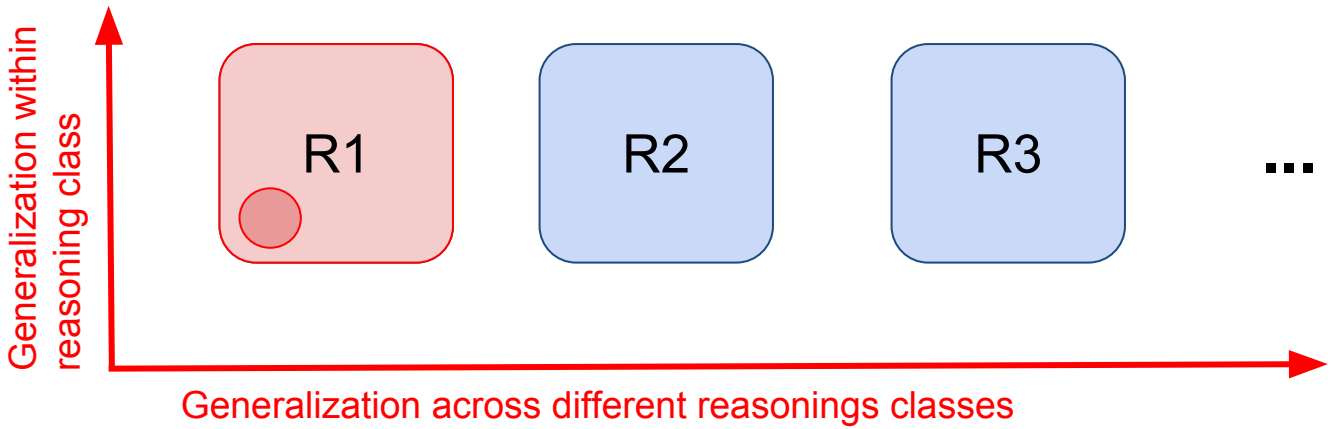
New Zealand shortest night

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

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

Proposal 2: QA "robustness" and "generalization"

Define equivalence class [R] to be the space of all (Q,A,K) tuples that are answerable using reasoning R.



- Two different axes of "reasoning" generalization:

Generalization within classes	If a solver answers (Q,A,K)∈[R], it should also be able to answer the rest of the questions in [R].
Generalization across classes	A solver should be able to cover more reasoning classes, upon seeing instances of classes {[R]}.

Proposal 2: QA “robustness” and “generalization”

Goal in this project:

- Create a reasoning-driven measure of robustness

Questions to Answer:

- Find ways to define the equivalence class
- Construct a playground with the definition of the equivalence class
- Evaluate existing SOTA systems with these measures.

Expected timeline

First idea:

- Initial implementation
- Reimplementing SemanticILP
- Reimplementing TableILP
- Unification or other extensions

~ 6 months

Second idea:

- Initial formalisms and pilot studies
- Creating the playground
- Evaluation

Conclusion

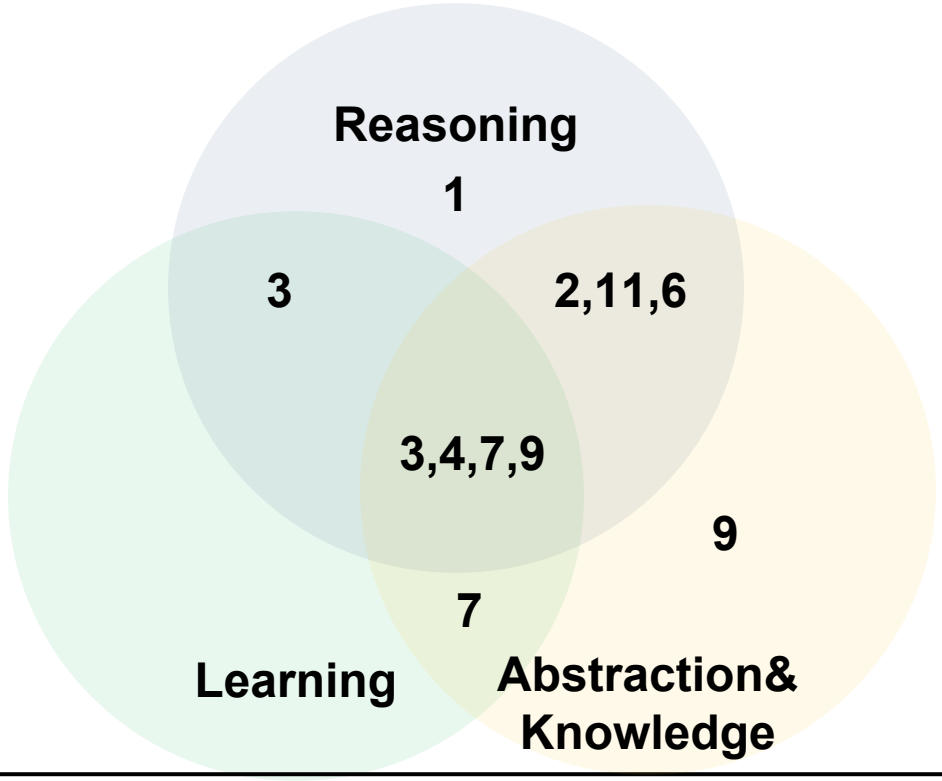
- Studying “reasoning” is a crucial element towards solving QA.
- We studied a few aspects of reasoning:
  - An abductive model, on top of *semantically-informed* representation.
  - What is *important* in questions
  - A playground for reasoning
- Next:
  - A programming language for reasoning over abstractions
  - A new notion “generalization”

And many issues remain open!

Thank you!

Questions?

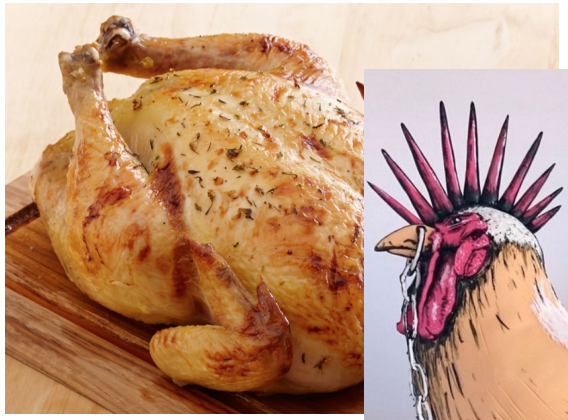
Cloud of relevant works



1. A challenge set for reading comprehension over multiple sentences, D. K., S. Chaturvedi, M. Roth, and D. Roth. 2018. (Submitted)
2. Question Answering As Reasoning on Semantic Abstractions, D.K, T. Khot, A. Sabharwal, and D. Roth, AAAI, 2018.
3. Learning What is Essential in Questions, D. K., T. Khot, A. Sabharwal, and D. Roth, CoNLL, 2017
4. Relational Learning and Feature Extraction by Querying over Heterogeneous Information Networks, P. Kordjamshidi, S. Singh, D.K, …, StarAI, 2017
5. Better call Saul: Flexible Programming for Learning and Inference in NLP, P. Kordjamshidi, D.K.,…, COLING, 2016.
6. Question Answering via Integer Programming over Semi-Structured Knowledge, D. K., T. Khot, A. Sabharwal, …, IJCAI, 2016
7. EDISON: Feature Extraction for NLP, Simplified, M. Sammons, C. Christodoulopoulos, P. Kordjamshidi, D.K., … , LREC, 2016
8. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions, P. Clark, O. Etzioni, D.K., … AAAI, 2016.
9. Illinois-Profiler: Knowledge Schemas at Scale, Z. Fei, D.K., H. Peng, H. Wu and D. Roth, Cognitum, 2015.
10. Solving Hard Co-reference Problems, H. Peng, D.K. and D. Roth, NAACL, 2015.
11. Flow of Semantics in Narratives, D.K, C. J.C. Burges, E. Renshaw, A. Pastusiak, Tech Report, August 2014.

What they missed: Language is context-dependent

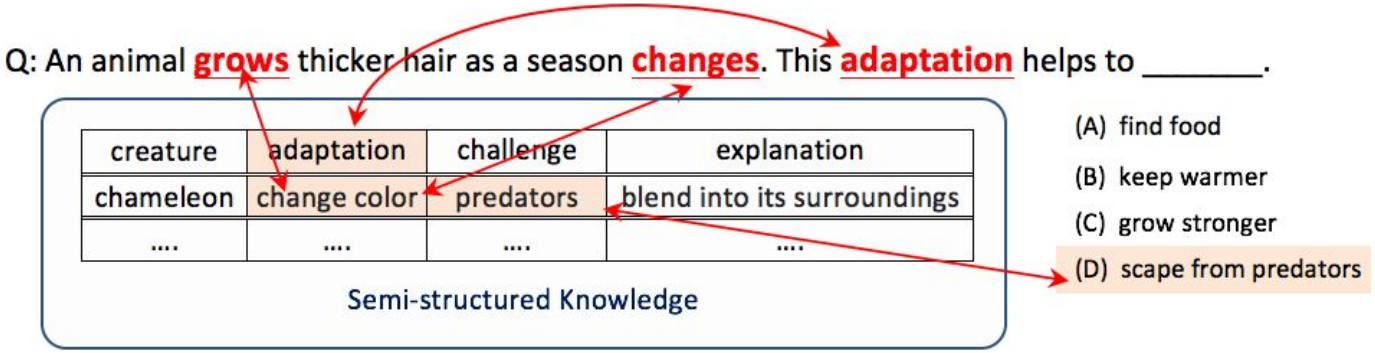
Chickens are ready



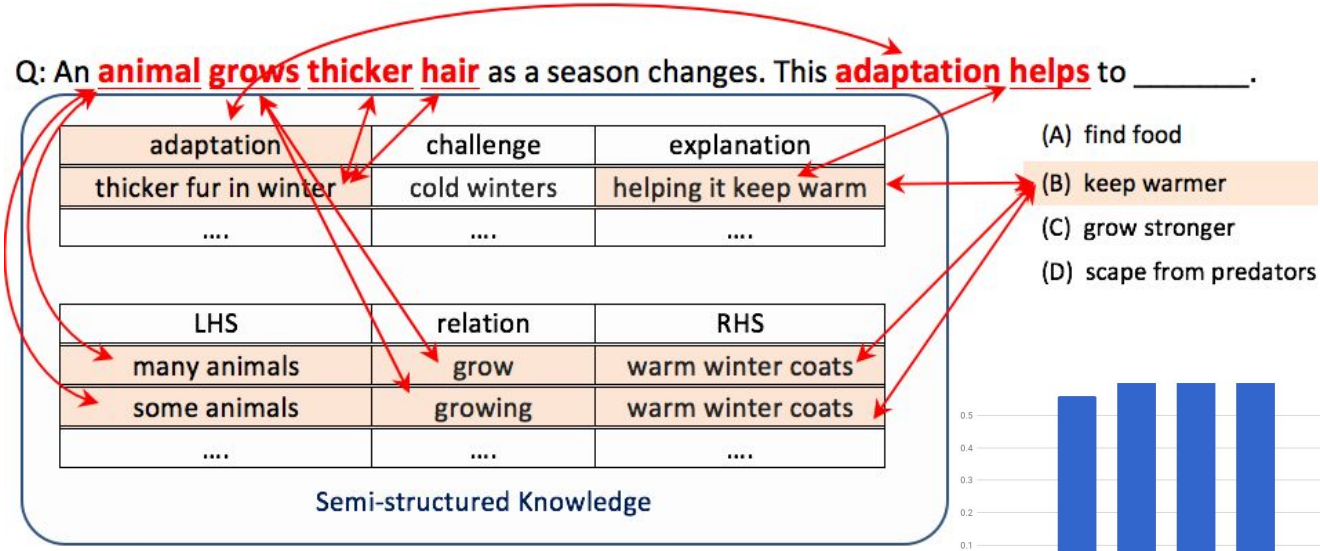
+ to eat

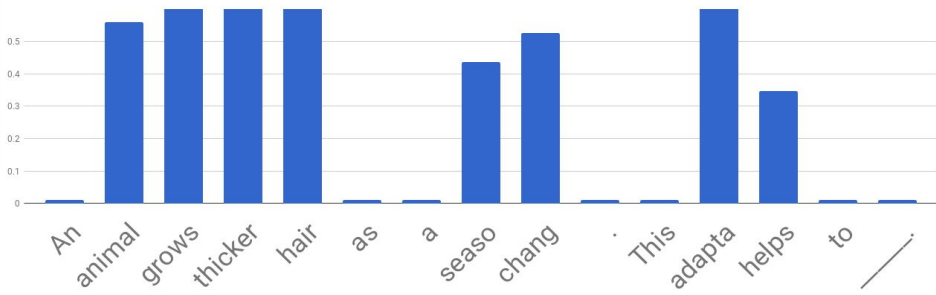
Lack of attention in TableILP

TableILP (Khashabi et al., 2016) does not recognize that "thicker hair" is an essential aspect of the question.



This problem is solved by augmenting the solver with essentiality scores:

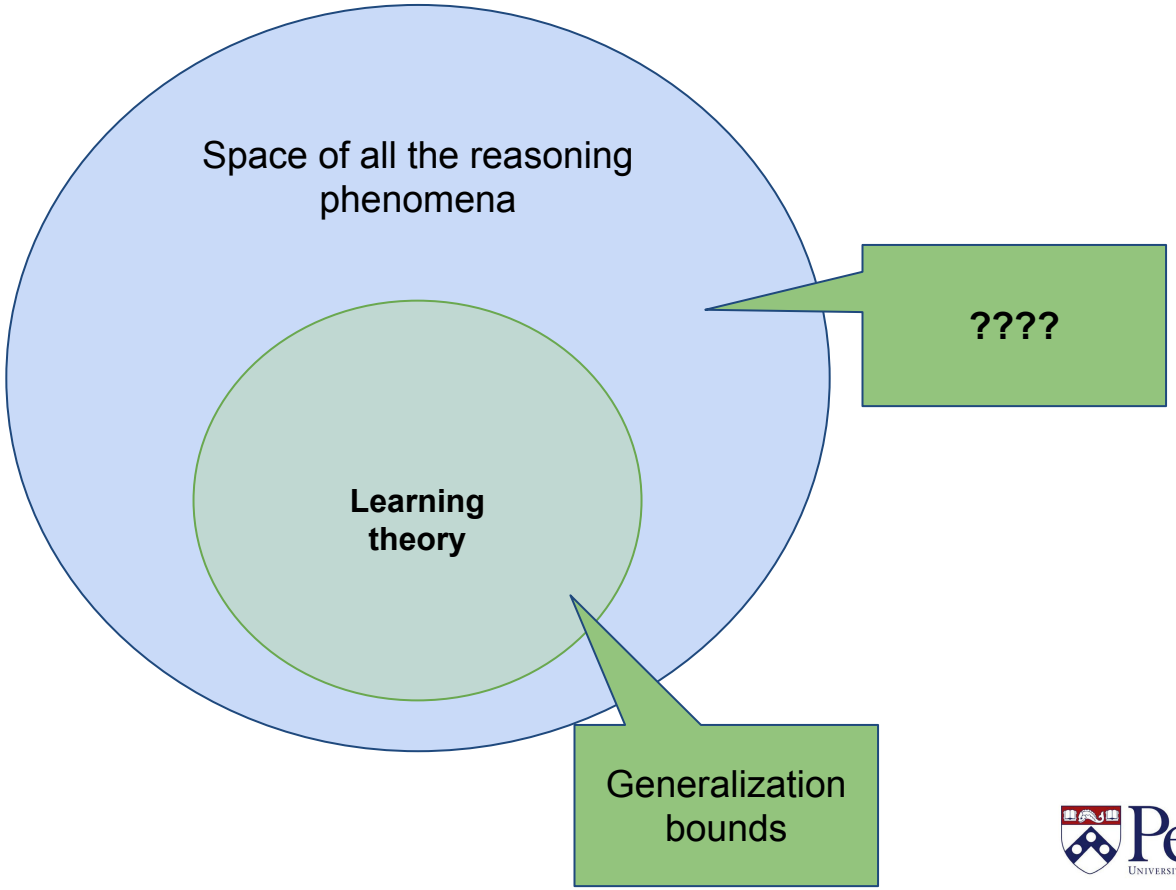




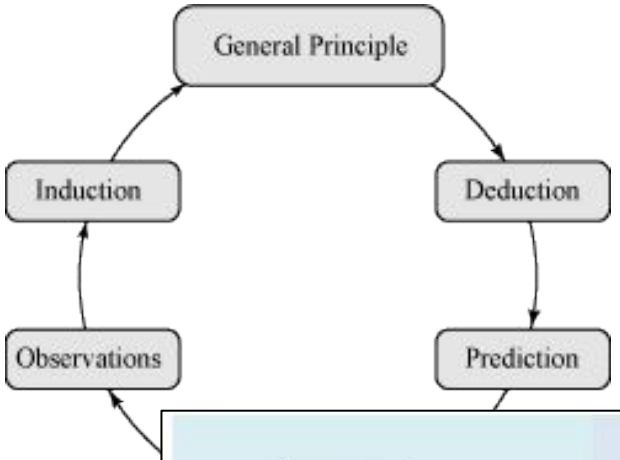
Proposal 2: QA “robustness” and “generalization”

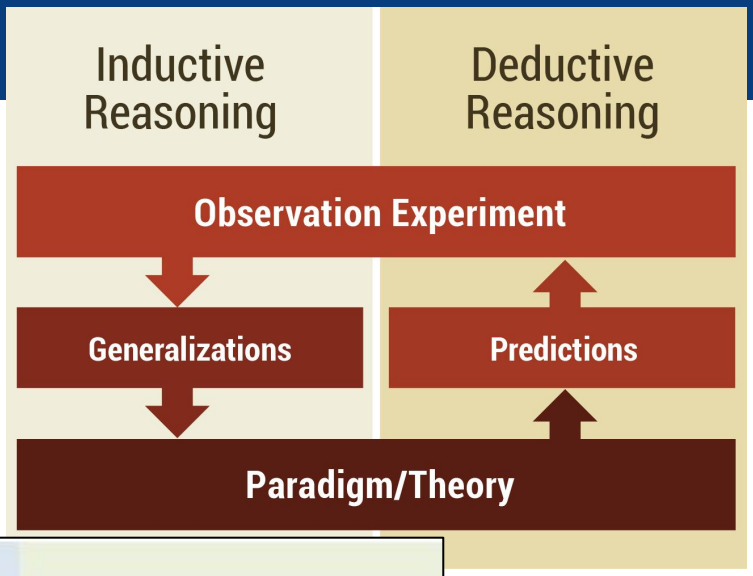
Goals in this project:

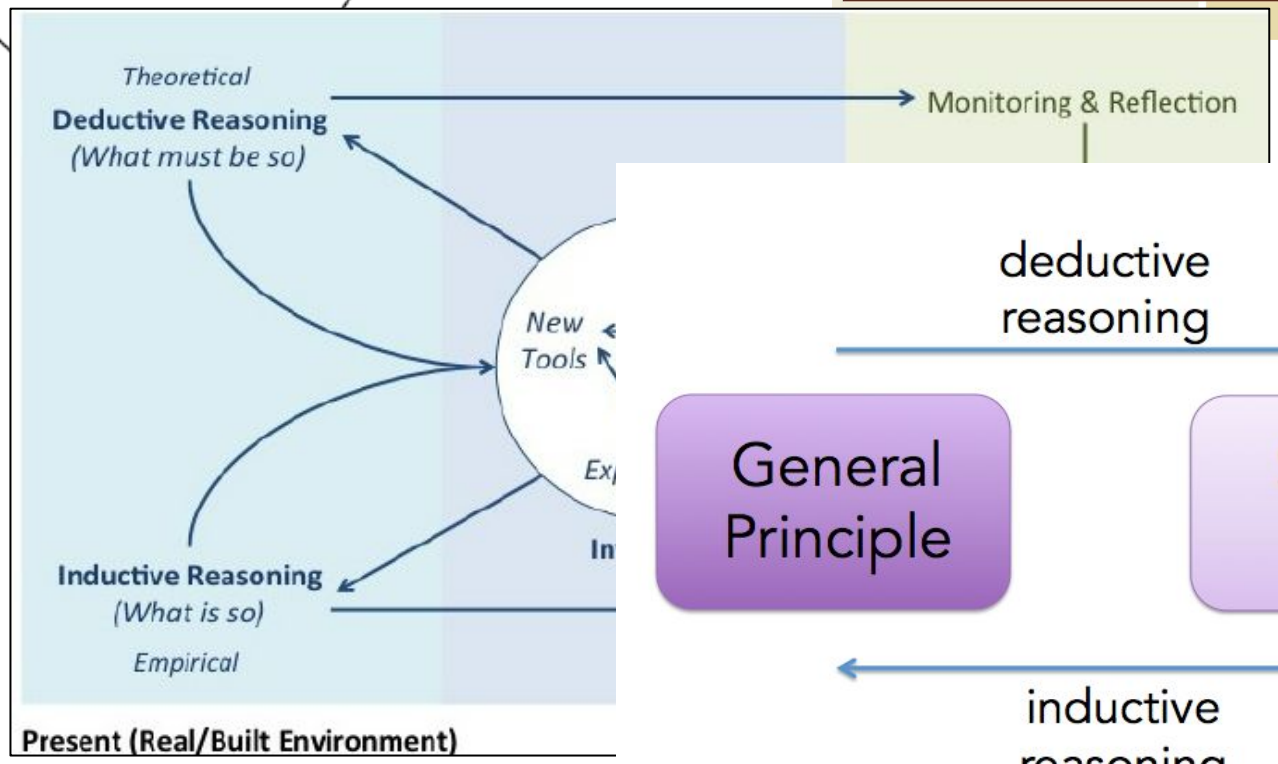
- Create a “linguistically-motivated” measure of robustness
- Steps towards measuring “genelization bound” beyond learning theory.



Reasoning in real life







EXTRA SLIDES

Many faces of reasoning

- **Abductive reasoning**

  The process of finding the best minimal explanation from a set of observations

  The grass is wet, …
  - It must have rained.
  - Someone has watered them

- **Inductive reasoning**

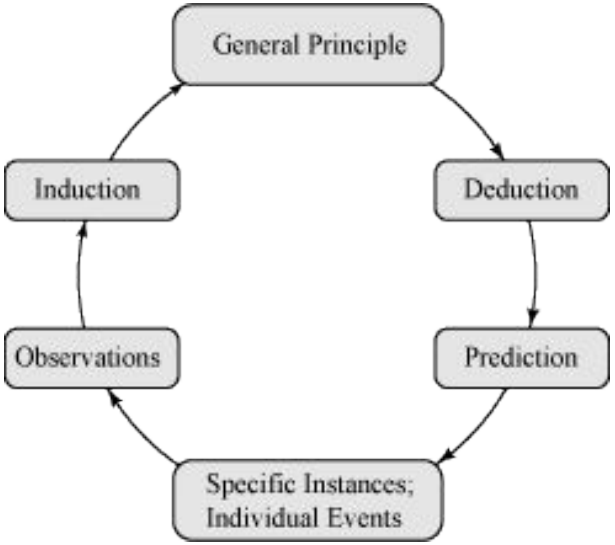
  The derivation of general principles from specific observations

  The grass has been wet every time it has rained.
  Thus, when it rains, the grass gets wet

- **Deductive reasoning**

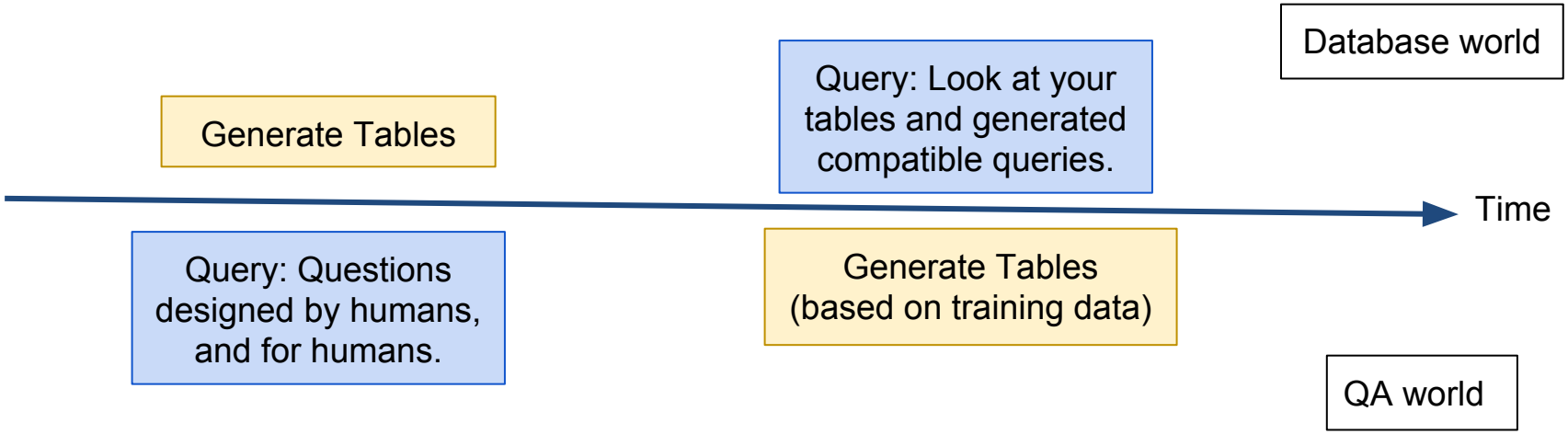
  Drawing conclusion from previous known facts and definitions

  The grass has been wet every time it has rained.
  Thus, when it rains, the grass gets wet



How does it compare to SQL query?

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

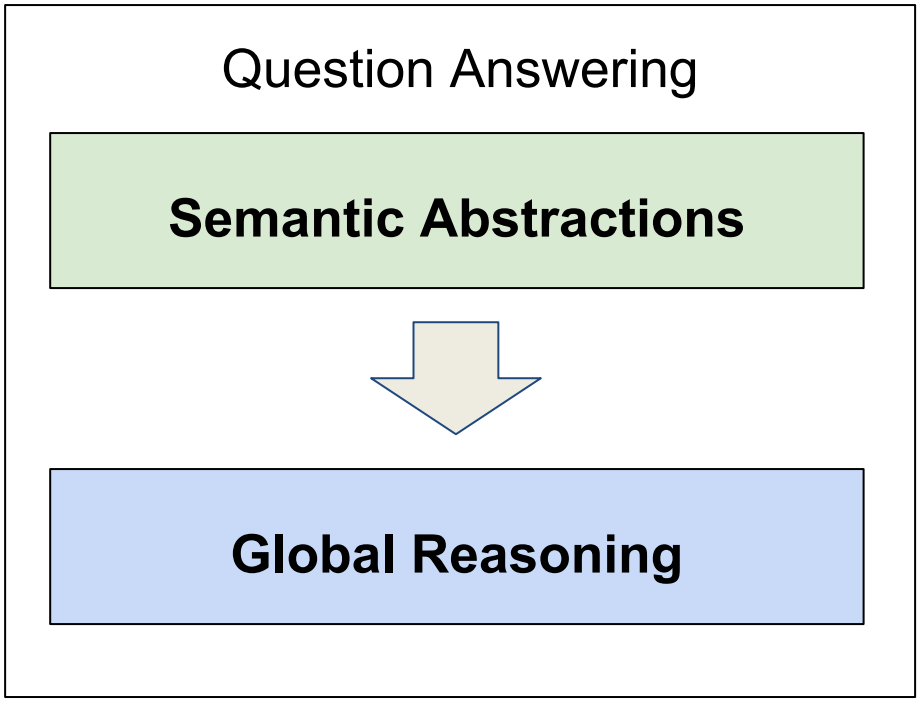


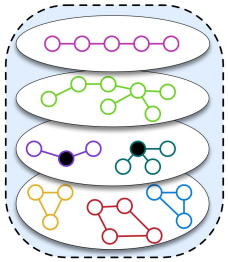
High-level view

Question Answering

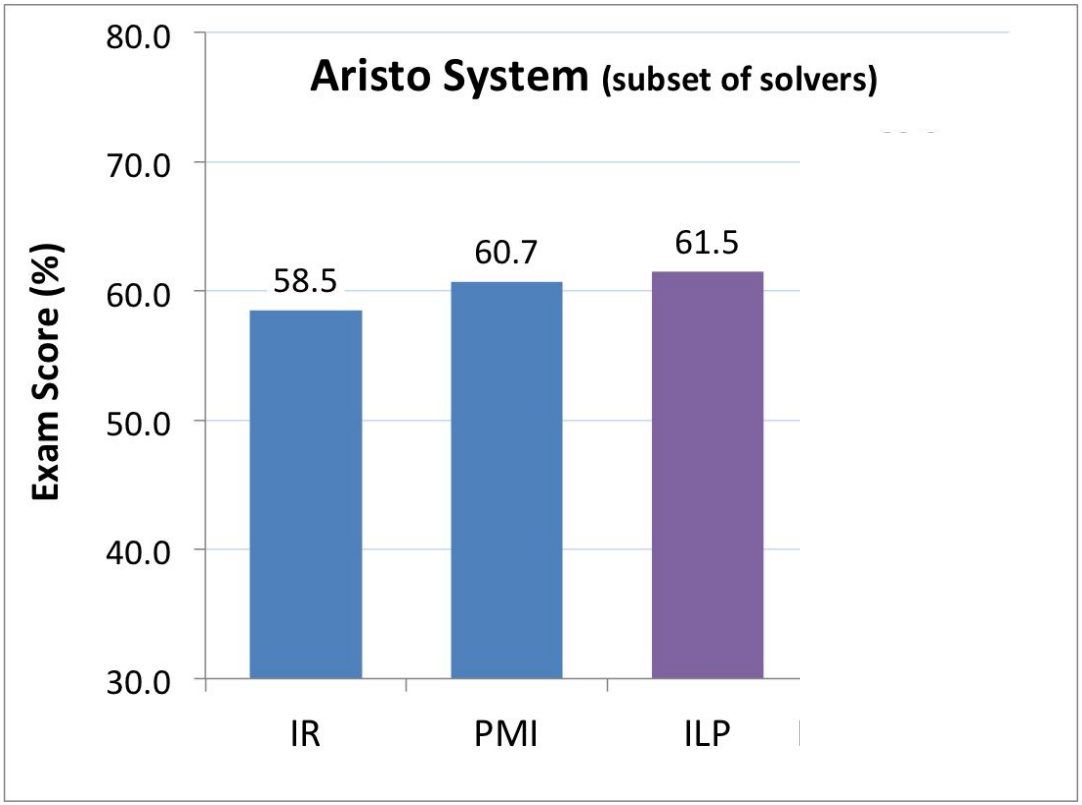
as Global Reasoning

over Semantic Abstractions





Results



Ensemble performs 8-10% higher than IR baselines

Simple logistic regression. Features: (Clark et al, AAAI-2016)

- 4 from each solver's score
- 11 from TableILP's support graph (#rows, weakest edge, …)

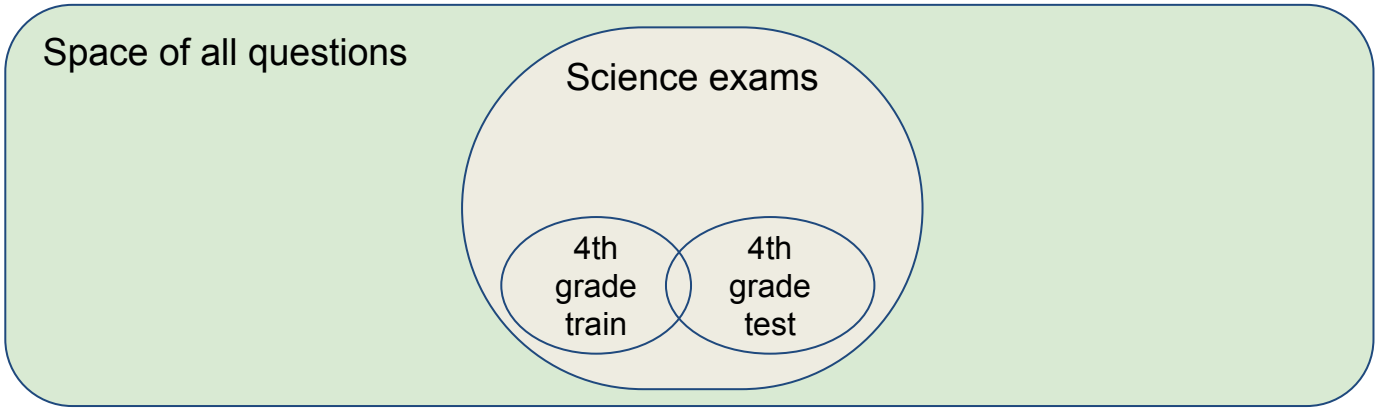
End-to-end systems: TableILP + ET

TableILP + ET

Employ a cascade system: *Questions unanswered by the first system are delegated to the second, and so on.*

Proposal 2: QA “robustness” and “generalization”

How QA systems are evaluated?

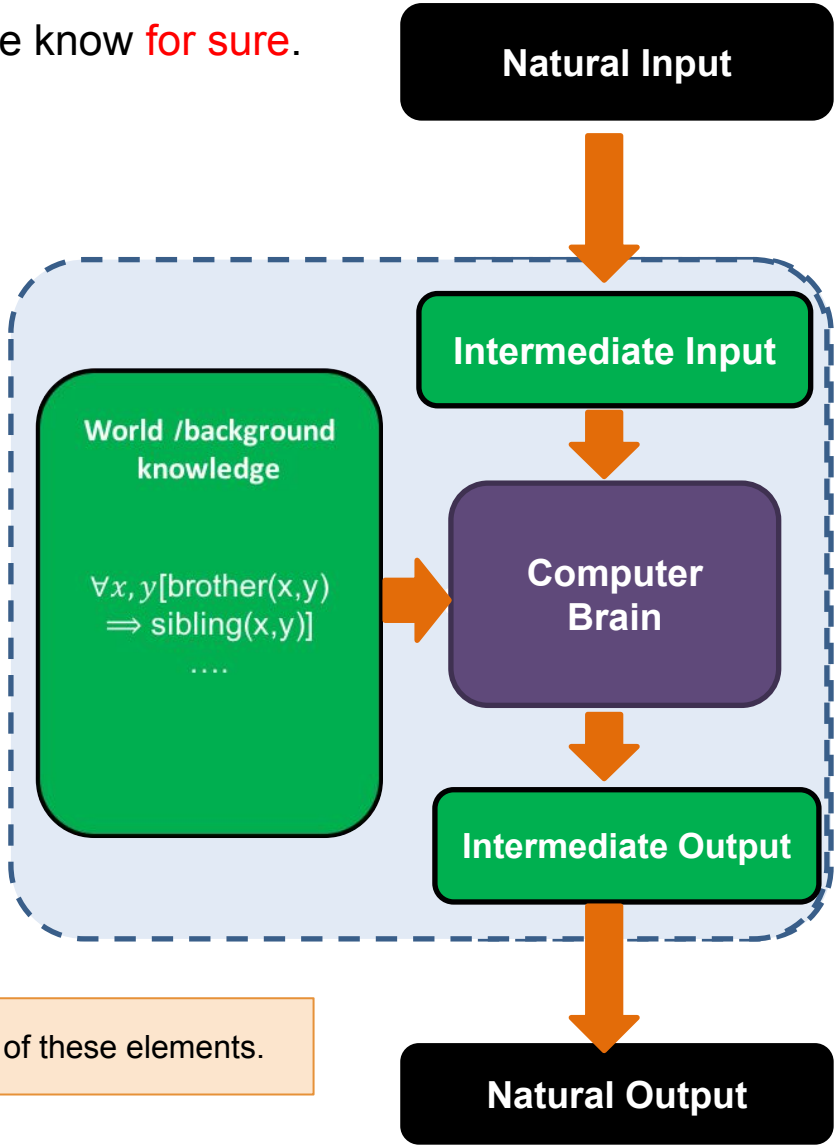


Question Answering problem

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



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

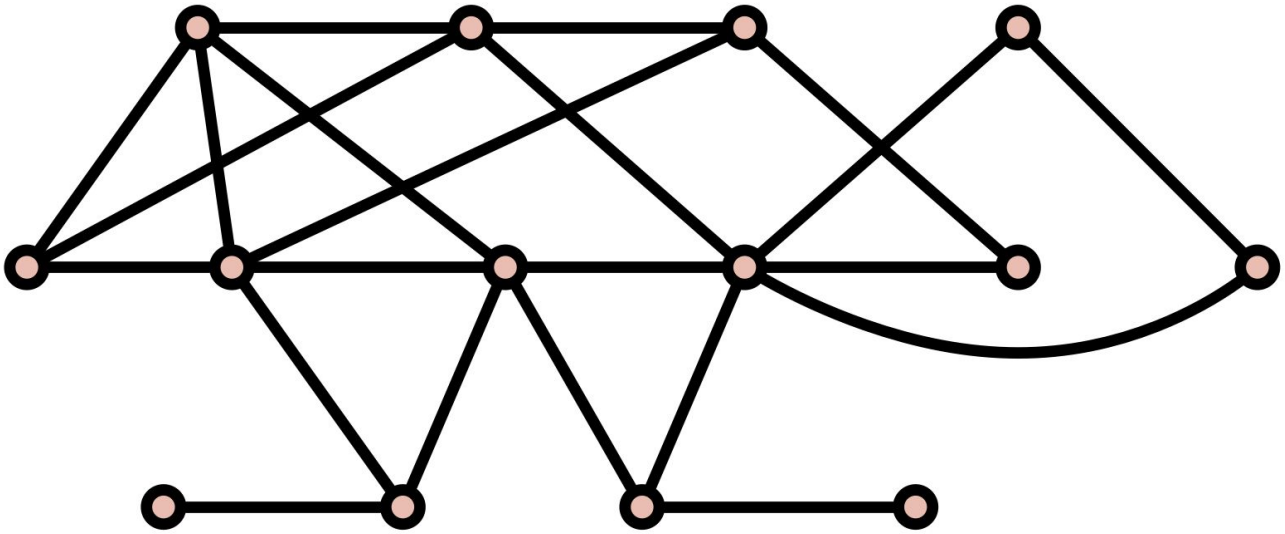
“Evaporation”

This does not entail isolation/independence of these elements.

Penn
University of Pennsylvania

Reasoning as max-likely explanation

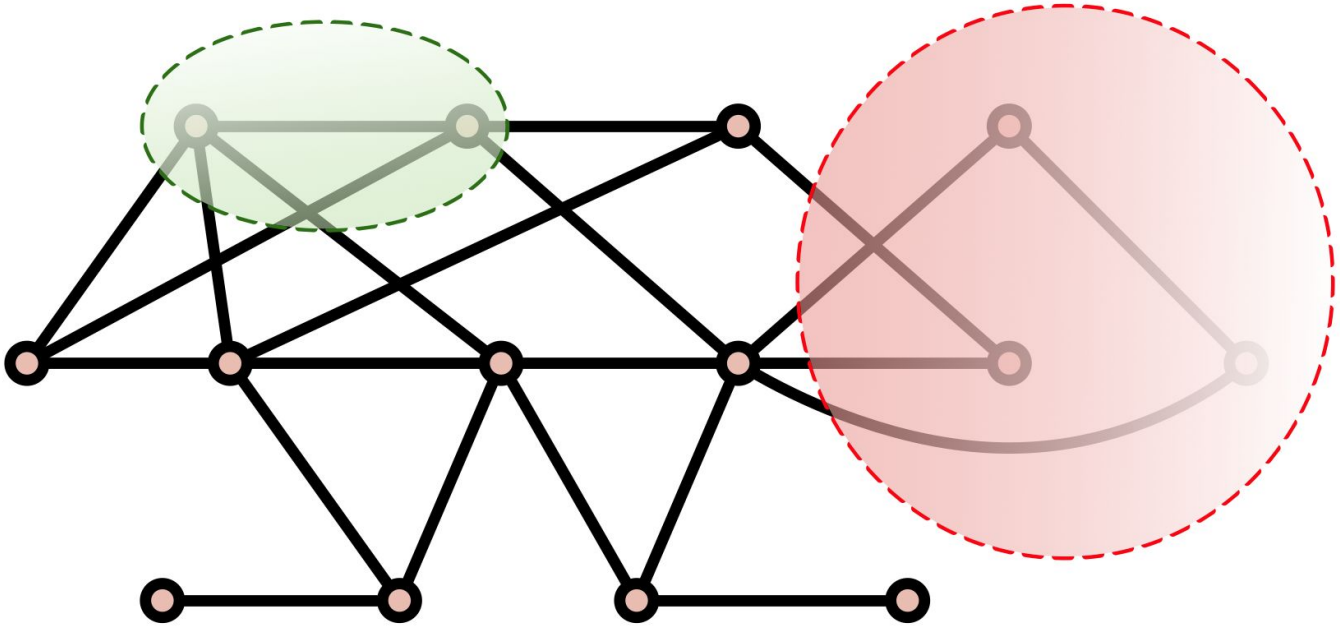
best simple explanation



$$G^* = \arg\max_{G \in \mathcal{G}} \text{score}(G)$$

Reasoning as max-likely explanation

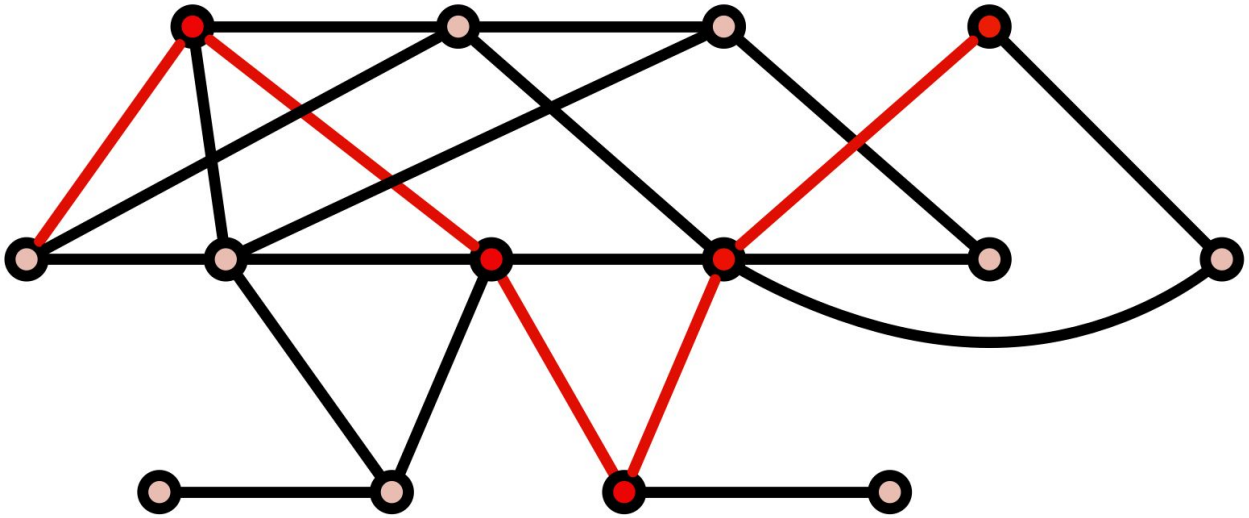
best simple explanation



$$G^* = \arg\max_{G \in \mathcal{G}} \text{score}(G)$$

Reasoning as max-likely explanation

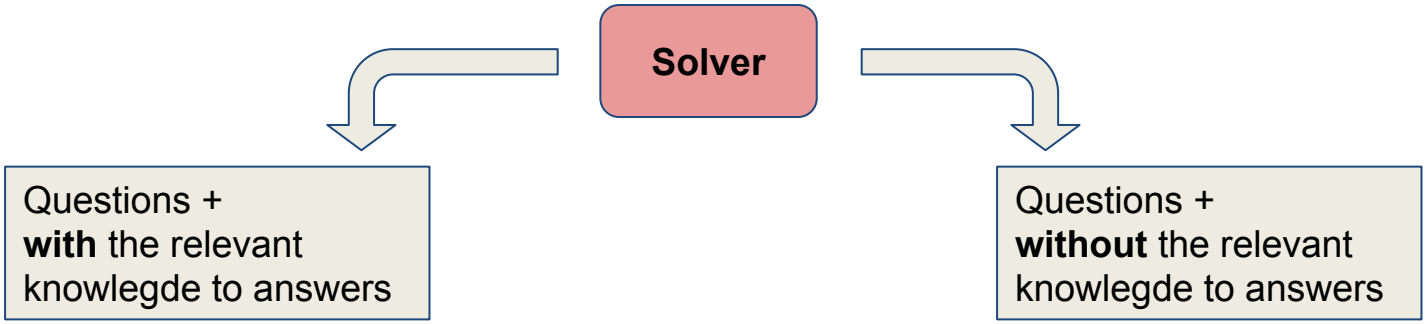
best simple explanation



“Don’t know” questions

- “don’t know” questions

How to know when we don’t know?

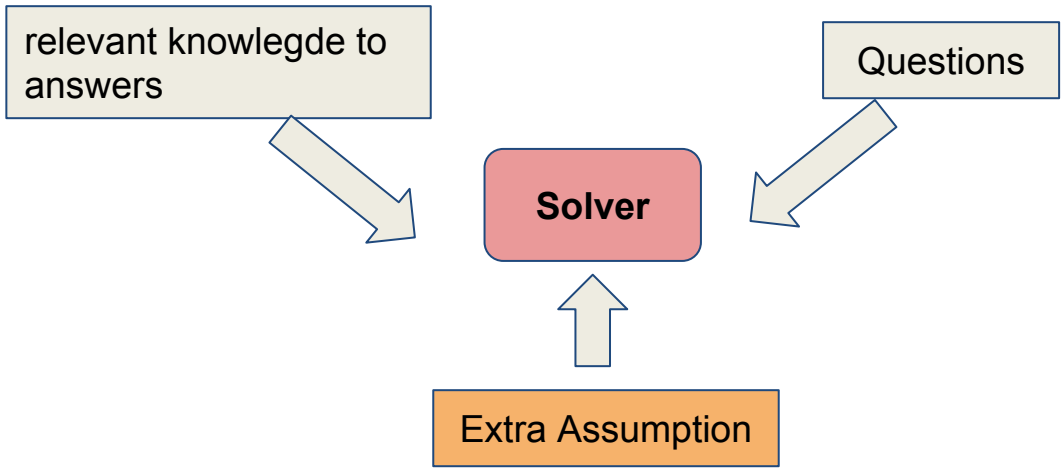


- Standardize the evaluation / propose a challenge
  - For given datasets create their “don’t know” dual.
- Evaluate existing approaches:
  - IR, NNs, Structured
- Looks for better approaches

Can we do better than thresholding?

“What if” questions

- “what if” questions



Extra assumptions could potentially change the answer to the questions.

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

“What if” questions (II)

New Zealand shortest night

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

New Zealand

What would save animals in a cold weather In ~~New York State~~?
(A) Size (B) Furry Skin (C) Long arms (D) Eating other animals

- Standardize the evaluation / propose a challenge
  - For given datasets create their “what if” extension.
- Evaluate existing approaches:
  - IR, NNs, Structured
- Looks for better approaches

How to represent “answers”

- The ultimate target should be getting only a question:
  - No candidates or reference text (hence, assumption on answers as substring of paragraph)

- There are two major paradigms used for QA:
  - (a) multi-choice questions (b) answer-substring-of-paragraph

- Both these methodologies have issues:
  - The assumption of “answer-substring-of-paragraph” is limited.
    - Not all questions have answers in a text as a contiguous substring.
      - For example implicit causes, or questions about a fictional scenarios.
    - Can be relaxed by using multiple [non-contiguous] spans.
  - Multiple-choice questions are limited, since it provides candidates
    - Having candidates is not trivial.
    - Too much extra information.

Current solutions to direct-answer QA

- **Generate candidates** and form a multi-choice queston

- Multiple reasons to have candidate generation:
  - **Engineering reason:** decouple process of answering question
  - **Conceptual level:** for many questions, it is not necessary to do the complete reasoning to generate candidates
    - For many questions candidate generation is easier than QA.
    - Need shallow reasoning for candidates and then carefully reading the paragraph for determining the answer (deep reasoning)

- Candidate generation have to be:
  - Diverse
  - Type compatible
  - High coverage
  - Domain adaptive

- Intrinsic evaluation
- Extrinsic (end-to-end) evaluation
  - Example: solve multiple-choice questions without answer-options.

Transferability in QA

QA systems suffer when moving to different “domains”

Forming the knowledge

- Genre
  - Vocabulary
  - Grammaticality
    - These two already covers the “Language” axis (English vs Spanish)
    - The temporal factor: language changes over time, e.g. for example twitter

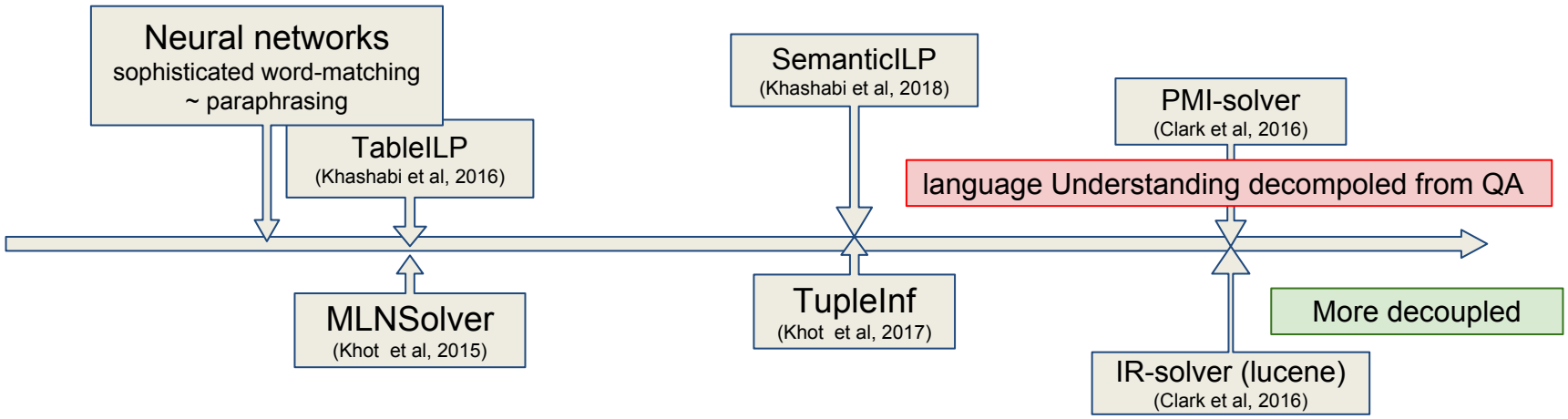
Reasoning

- Label space: do we want to multiple choice (single-correct or unspecified), or substring of paragraph, or direct–answer, etc
- Reasoning type (causal reasoning, temporal reasoning, …)

- There is no clear definition of QA system being “domain adaptive”
- Often times “reasoning type” is conflated with “text genre”

Language understanding + QA

How much do system decouple their language understanding from QA?



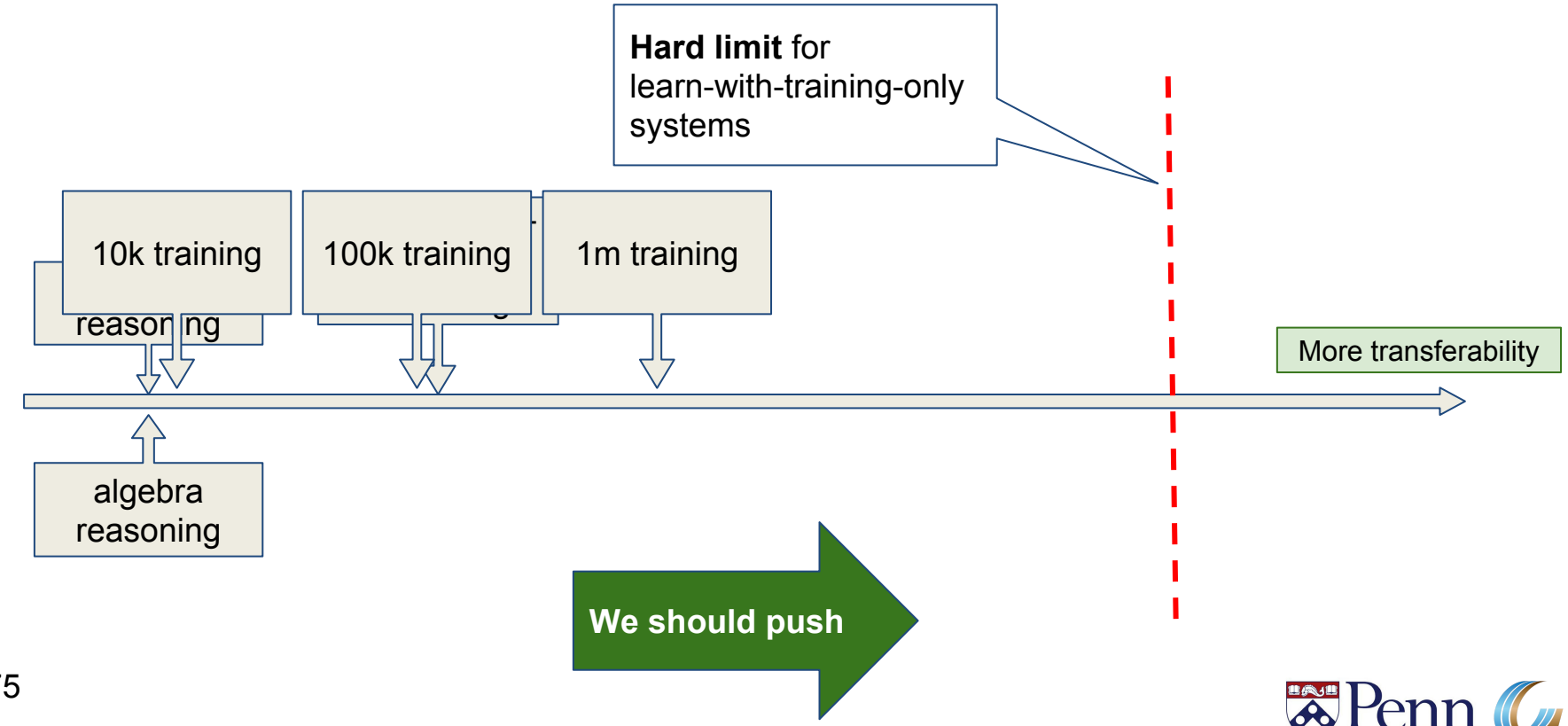
Transferability

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.



Pushing the transferability

- Solvers working in multiple “datasets”
- Learning with a few instances; e.g.
  - A solver is “trained” for a domain
  - Given “a few” instaces from another domain, it should adapt itself the new domain.
- Easy way for incremental supervision; e.g. by
  - Giving “instructions” in the form of knowledge
  - Showing instances of “bad” or “good” reasonings

2. Domains and Adaptation for Reasoning

- Variability in reasoning:
  - Assume that “pipeline” issues are resolved.
  - Can we do something in terms of reasoning?
    - In SemanticILP, reasoning is everything representable as some (undirected?) alignment.
    - Suggestion rather than defining reasoning we can discuss the contribution of linguistic phenomena.
  - One can consider two scenarios: (1) completeness in reasoning (2) or not.
  - In (1) the issue is mostly finding the balance between reasoning methods.
  - In (2) the challenge is about “learning new reasoning”.
    - This sounds more challenging than the first aspect we discussed.

- Lack of metric:
  - Creating explicit measure of domain-similarity could be helpful. Potentially on a dataset.

3. A high-level language to define reasoning

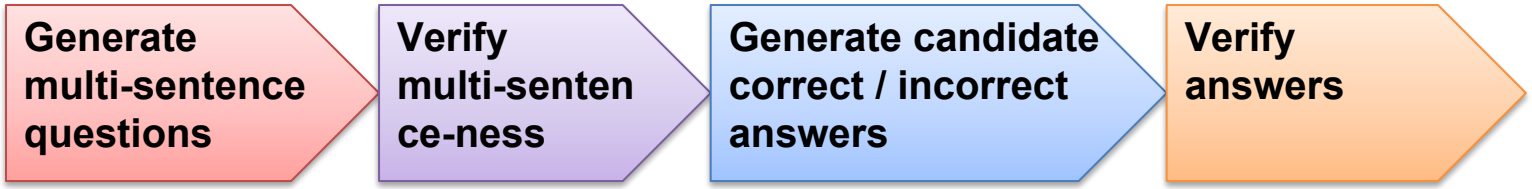
- 3 papers on ILP-based QA and each thousands of lines
  - Can we simplify the process of defining reasoning?
  - Suppose everything is presentable with TextAnnotation views
  - Assumption: the reasoning graph is,
    - connected,
    - no dead-end in paragraph.
  - A use defines:
    - What views from question can be aligned to what views of paragraph
    - What views of paragraph can be aligned to what views of answer
    - What views of paragraph can be aligned to each other
      - And how their edge weighted
    - For each view type, define the constraints:
      - E.g. for predicate-argument type you have to use predicate and at least an argument
    - Pre-defined global weights/constraints:
      - Max/min number of constituents can be used in Question, Paragraph, each Ans
      - Max/min number of edges connected to each constituent, in Question, Paragraph, and each Ans
- Claim is:
  - this is an easy way to define reasoning, and it subsumes many existing definitions (show how simply you can define SemanticILP)
  - Create SemanticILP + TupleILP + TableILP
  - Or show that you can extend it:
    - E.g. SRL alignments inside table cells.

3.1 Beyond simplified definition

- The language can be further simplified for people who don't have much understanding about "graphs", etc.
  - Essentially the users should be able to define interesting patterns and uninteresting patterns, and the system should be able to do infer based upon that.
  - This can also turn into an inductive system.

1.Measuring reasoning capability of QA systems

- What is reasoning?
  - We don't define it, but we assume that it often involves multiple sentences.
- Suggested solution: create a dataset
  - A paragraph, with a set of multi-choice questions



- Currently:
  - 50 passage processed
  - generated 300 questions
  - Cost ~140$
  - For 30k questions we have to spend 13.8k$
- What interesting analysis can we do on this?

5. What does it take to beat a neural network?

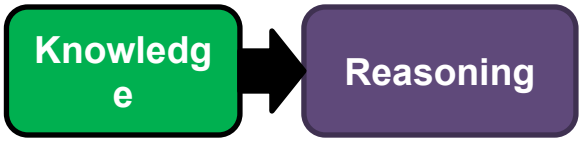
- Create a linear system and beat existing neural network reading comprehension systems.

Slightly different paradigm for DA QA

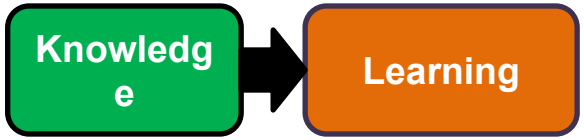
- DA questions shouldn't be limited to spans of a given paragraph
  - Although they should be consistent with it.
- The main issue is evaluation

Decision Making: Interface between *learning* and *reasoning*

- Reasoning only
  - Symbolic reasoners accessing knowledge directly
  - General Problem Solver (Simion&Newell,1956)



- Learning + Knowledge
  - End-to-end learning, e.g. BiDaF (Seo et al, 2016)

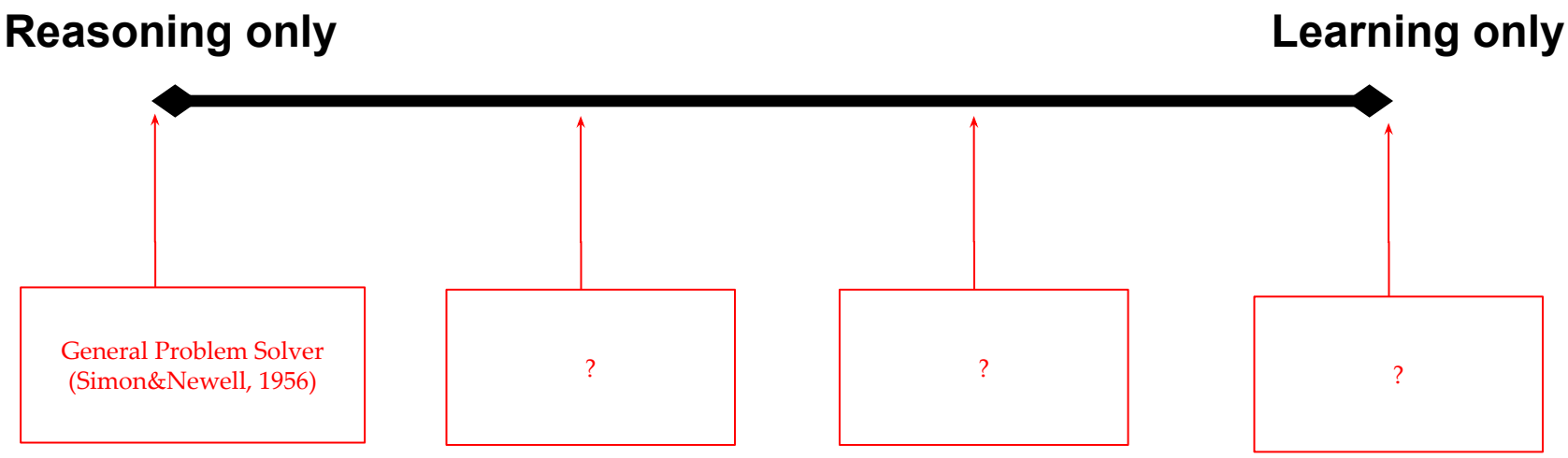


- Reasoning + Learning + Knowledge:
  - *Learning* as soft-accessability to knowledge for Reasoning; e.g.TableILP (Khashabi etl al, 2016)
  - *Learning* as decision-maker, followed by reasoning (Yih and Roth, 2004)
  - Learning to extract Reason-ble structure from input; e.g. semantic parsers (Pradhan et el, 2004)

2. Domains and Domain adaptability

- Define domain:
  - Genre
    - Vocabulary
    - Grammaticality
      - This already covers the “Language” axis (English vs Spanish)
      - The temporal factor: language changes over time, e.g. for example twitter
  - Label space
  - Reasoning type (causal reasoning, temporal reasoning, …)
- To what extent these axes are independent?
- Tasks:
  - NER, Mention, QA
- Analysis:
  - Using these metrics, come up with explicit measure of domain-similarity
  - Can you optimize for domain-similarity?
    - Modify the training data (across different axis) and show that you can match to a certain measure …
  - Take one important result from the past 10 years and show that it is using each axis
    - Then fix the issue in a principles way
  - How do word-vectors play out here? (vocabulary, label-space, etc)
  - After defining/studying “domain”, we have to define “generalization”

Learning vs Reasoning spectrum



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

Example

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}*

Example

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}*

Simple "lookup" based on proximity to question words, answer type

- Basic word overlap suffices
- Neural methods (e.g., BiDAF) excel at

Example, Rephrased

P: 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.

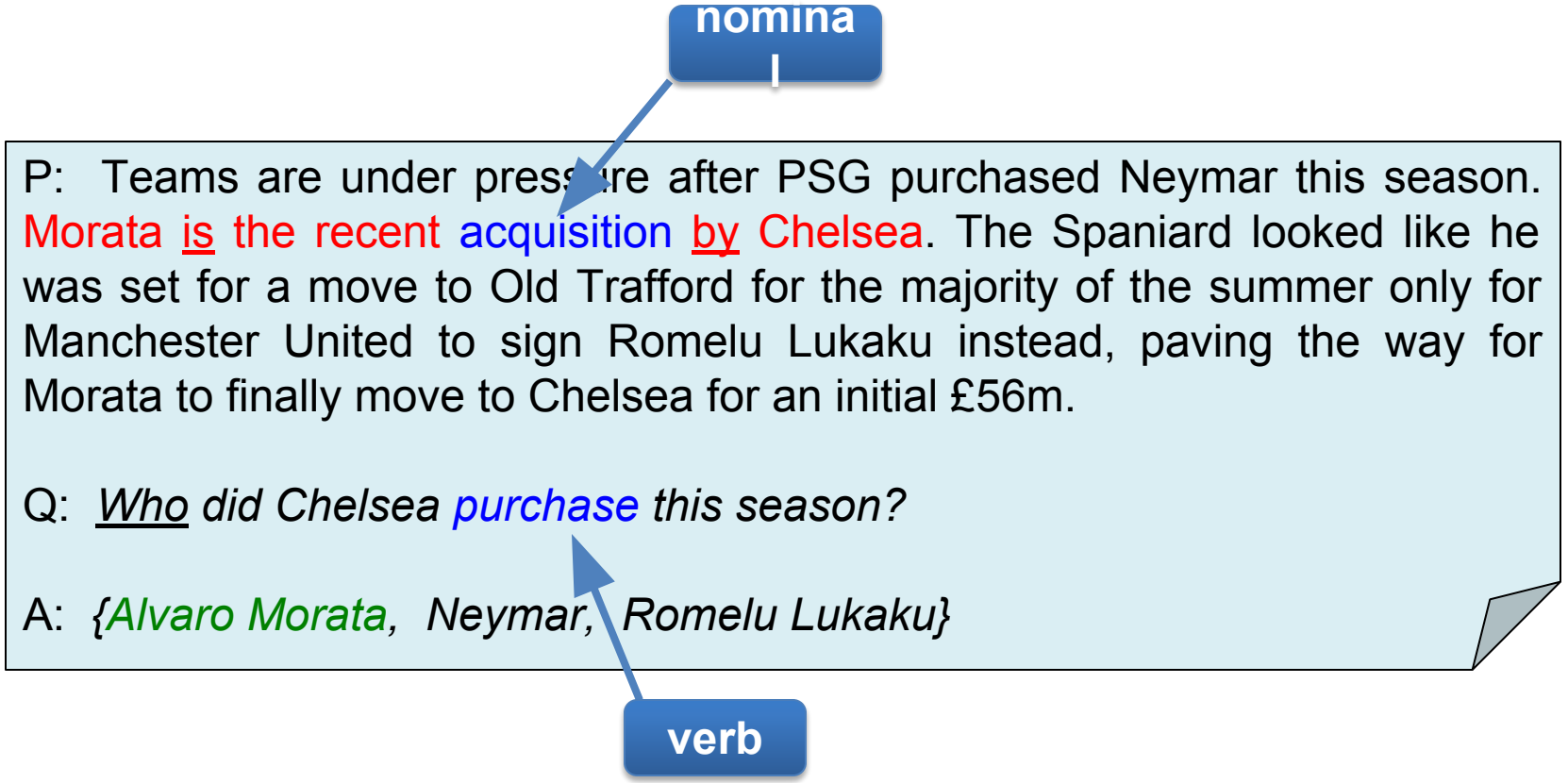
Q: *Who did Chelsea purchase this season?*

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

Simple rewording can confuse solvers

- E.g., BiDAF outputs *"Neymar this season. Morata"*

Example, Rephrased



Linguistic understanding can help!

- Verbs and their nominalization
- Domain agnostic => can use pre-trained NLP modules

Example, Rephrasing #2

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.

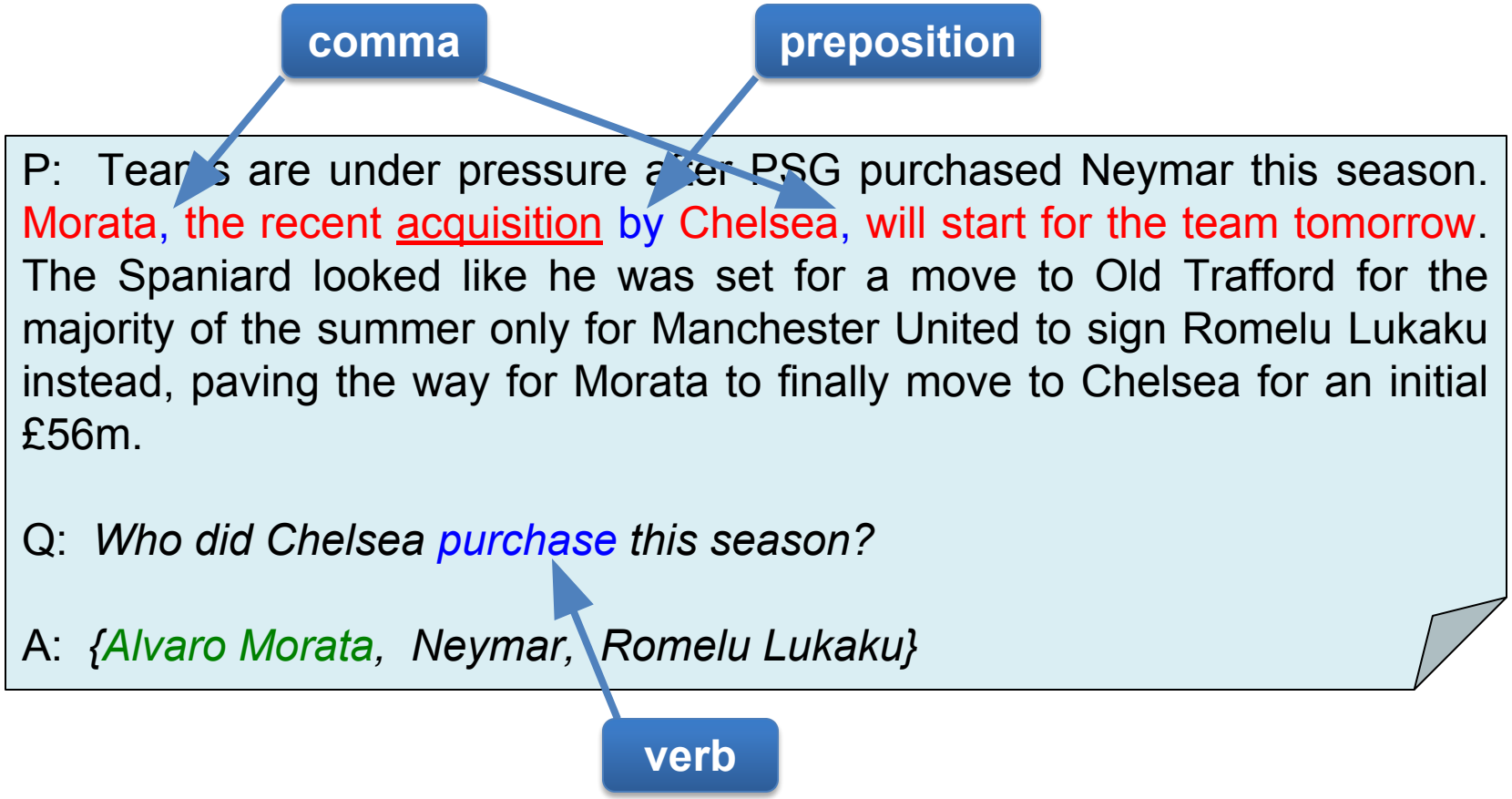
Q: *Who did Chelsea purchase this season?*

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

Simple rewording can confuse solvers

- E.g., BiDAF outputs *"Neymar"*

Example, Rephrasing #2



Linguistic understanding can help!

- Verbs, preposition, punctuation
- Domain agnostic => can use pre-trained NLP modules

SemanticILP

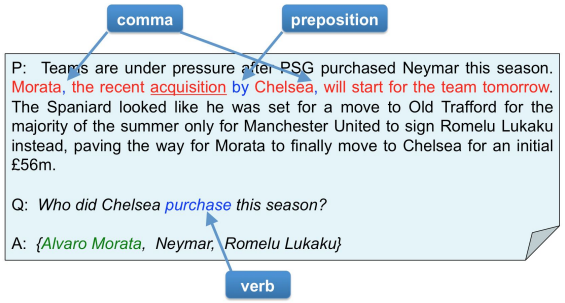
1. Create a collection of semantic abstractions of text (all of Q, A, P)
   - Use off-the-shelf, pre-trained NLP modules
   - Multiple views for a more complete semantic understanding
     - SRL frames (verbs, prepositions, comma), dependency parse, coreference sets, lexical similarity links, raw text sequence

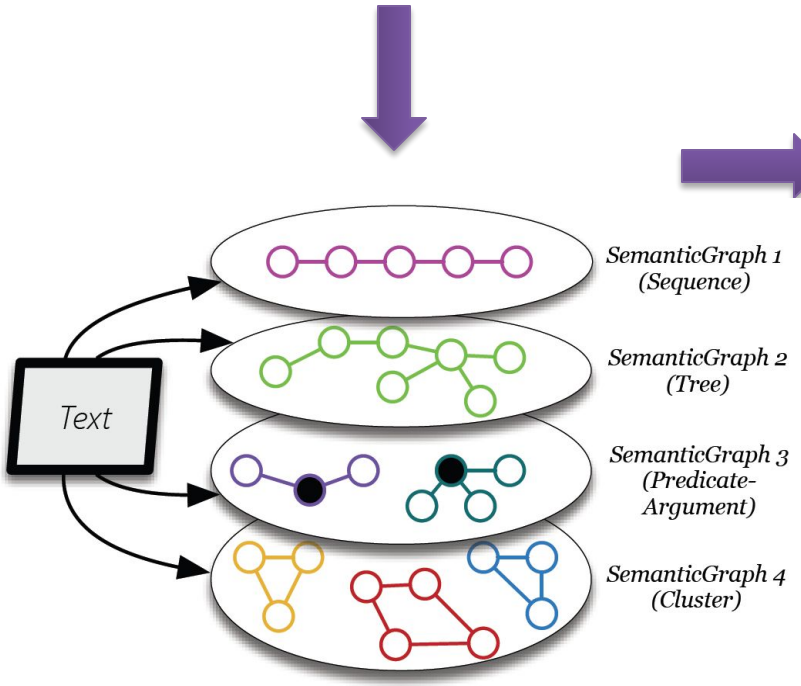
2. Create a unified representation as a family of graphs
   - PredArg graphs, trees, clusters, sequences
   - Connected via textual similarity links

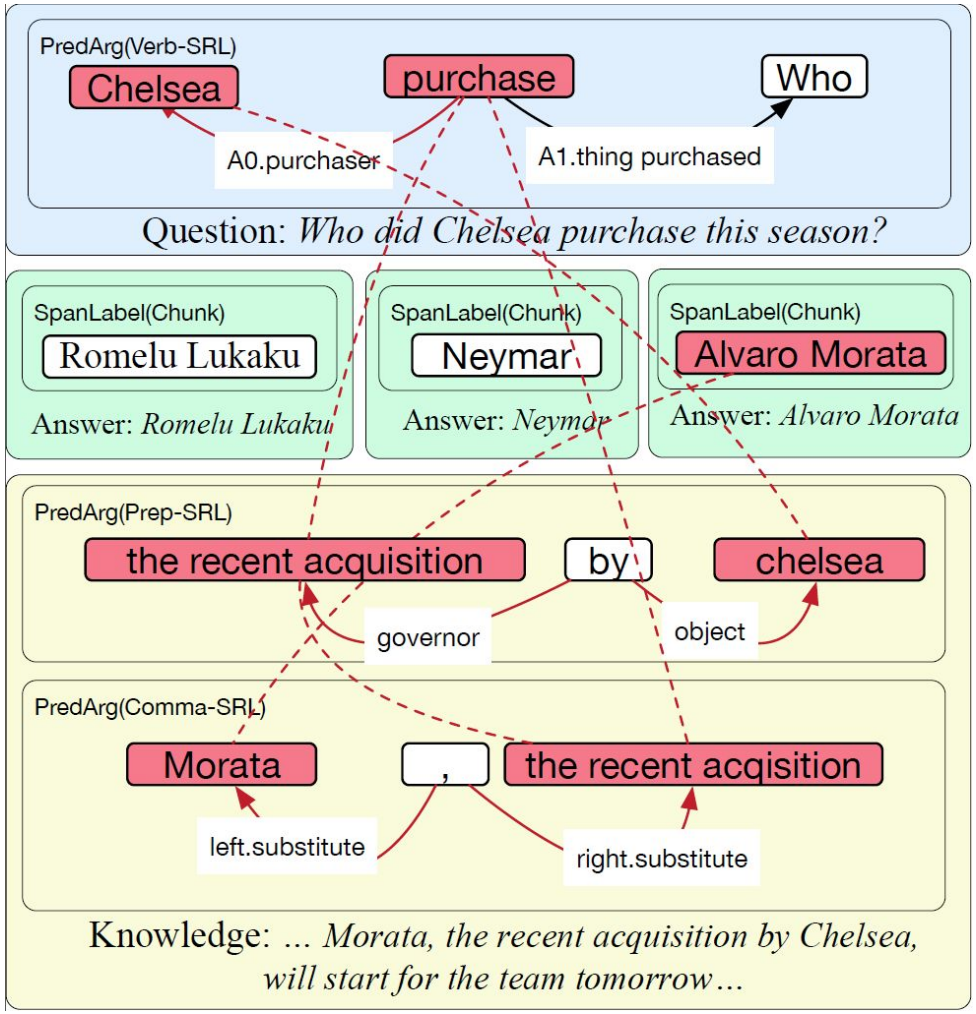
3. Translate QA into a search for an optimal subgraph
   - Incorporate global and local constraints and preferences
   - Capture what's a valid reasoning, what's preferred

4. Formulate as Integer Linear Program (ILP) optimization
   - Solution points to the best supported answer

SemanticILP: Example







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

Dataset	**BiDAF**	**BiDAF tuned**	**IR**	**TupleInf** [ACL-2017]	**SemanticILP** (linear comb. of components)
Regents 4th	56.3	53.1	59.3	61.4	67.6
Public 4th	50.7	57.4	54.9	56.1	59.7
Regents 8th	53.5	62.8	64.2	61.3	66.0
Public 8th	47.7	51.9	52.8	51.6	54.8

(exam scores, shown as a percentage)

Results #2: ProcessBank [EMNLP-2014]

- Input: **Biology question** Q with 2 answer options A, paragraph P

Dataset	BiDAF	BiDAF tuned	IR	SyntProx Baseline*	ProRead* (structural supervision)	SemanticILP (linear comb. of components)
Process Bank**	58.7	61.3	63.8	61.9	68.1	68.6

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

* Berant et al. [EMNLP-2014]

** ~70% of the original dataset; true/false and temporal questions currently out of scope

SemanticILP: Summary

- First QA system to combine multiple semantic abstractions
  for a more complete understanding of text
- State-of-the-art results on two datasets with different characteristics
- Extensible architecture
  - Expand semantics via new NLP modules (e.g., QA-SRL, temporal)
  - Expand to different kinds of reasoning (e.g., causal sequences)
- A promising knowledge representation formalism
  - Unit of knowledge => paragraph P
  - Semantics => graph representation of abstractions of P

Key Challenges and Solutions

A. Textual knowledge is expressed in a variety of linguistic forms
   - No single knowledge representation (e.g., Open IE tuples) suffices

   => *Broader coverage via multiple kinds of NLP modules*

B. NLP systems for these are noisy
   - SRL, coref, shallow parsers, chunkers, …

   => *Robustness via multiple modules of the same kind*

C. Combining information from multiple NLP modules is non-trivial

   => *Global consistency via global ILP optimization*

2. Domains and Domain adaptability

- Define domain:
  - Genre
    - Vocabulary
    - Grammaticality
      - This already covers the “Language” axis (English vs Spanish)
      - The temporal factor: language changes over time, e.g. for example twitter
  - Label space
  - Reasoning type (causal reasoning, temporal reasoning, …)
- To what extent these axes are independent?
- Tasks:
  - NER, Mention, QA
- Analysis:
  - Using these metrics, come up with explicit measure of domain-similarity
  - Can you optimize for domain-similarity?
    - Modify the training data (across different axis) and show that you can match to a certain measure …
  - Take one important result from the past 10 years and show that it is using each axis
    - Then fix the issue in a principles way
  - How do word-vectors play out here? (vocabulary, label-space, etc)
  - After defining/studying “domain”, we have to define “generalization”

Friday (Nov, 10)

Short term

- System's paper (Dec 8th, CPAIOR)
  - Gist: simple definitions SemanticILP
- Canidate generation (Jan 10th, NAACL short)
  - Goal: being able to answer [subset of] aristo 4th grade questions without candidate answers

Long term

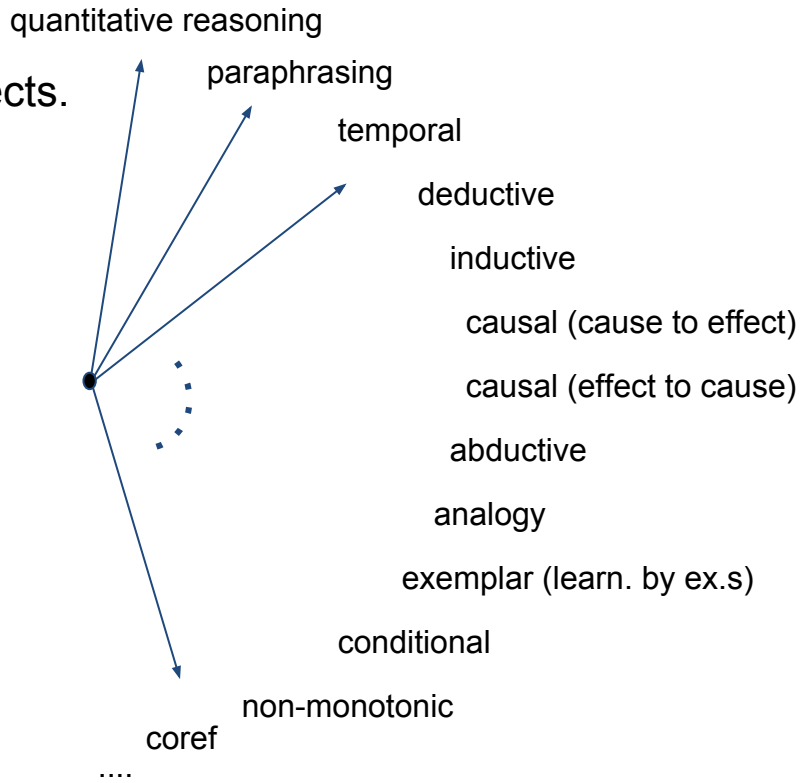
- Can we create a combination with learning-only systems?
- Some reasonings are missing:
  - Learning by examples (~induction)
  - Conditional (what if ….)
  - Yes/No questions
  - Don't know questions
  - Temporal events
  - Quantitative and / or Algebra
  - Analogy
  -

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.



AI Goal:
Towards natural language understanding.

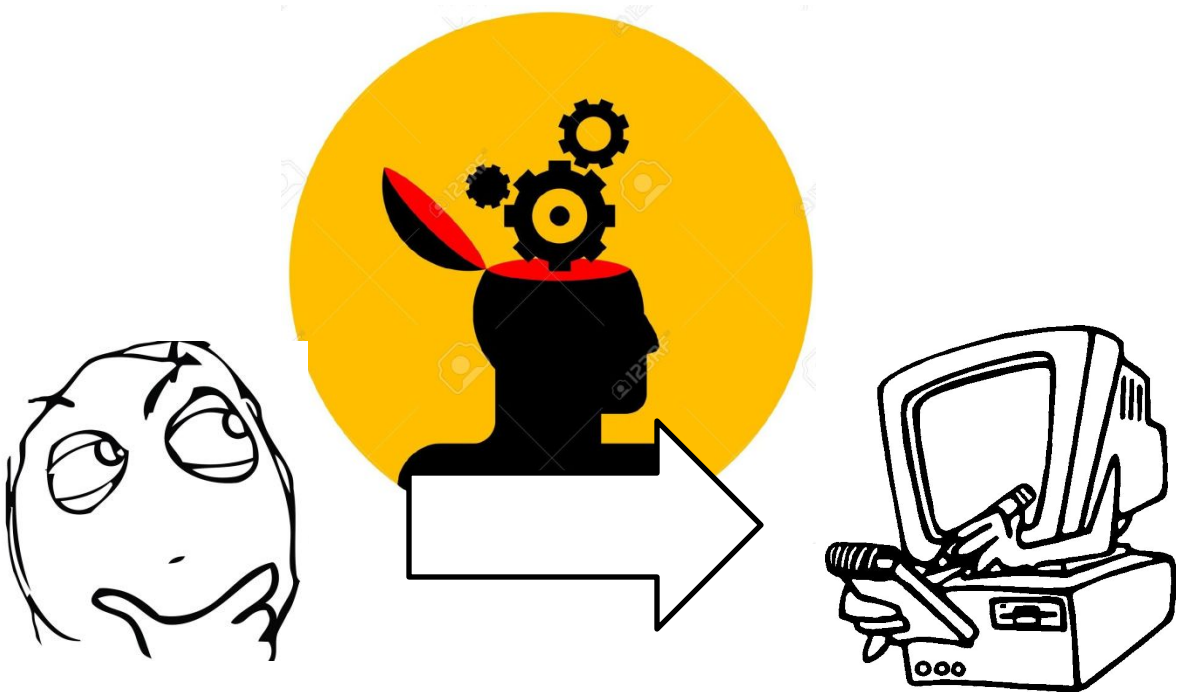
How to measure the progress?

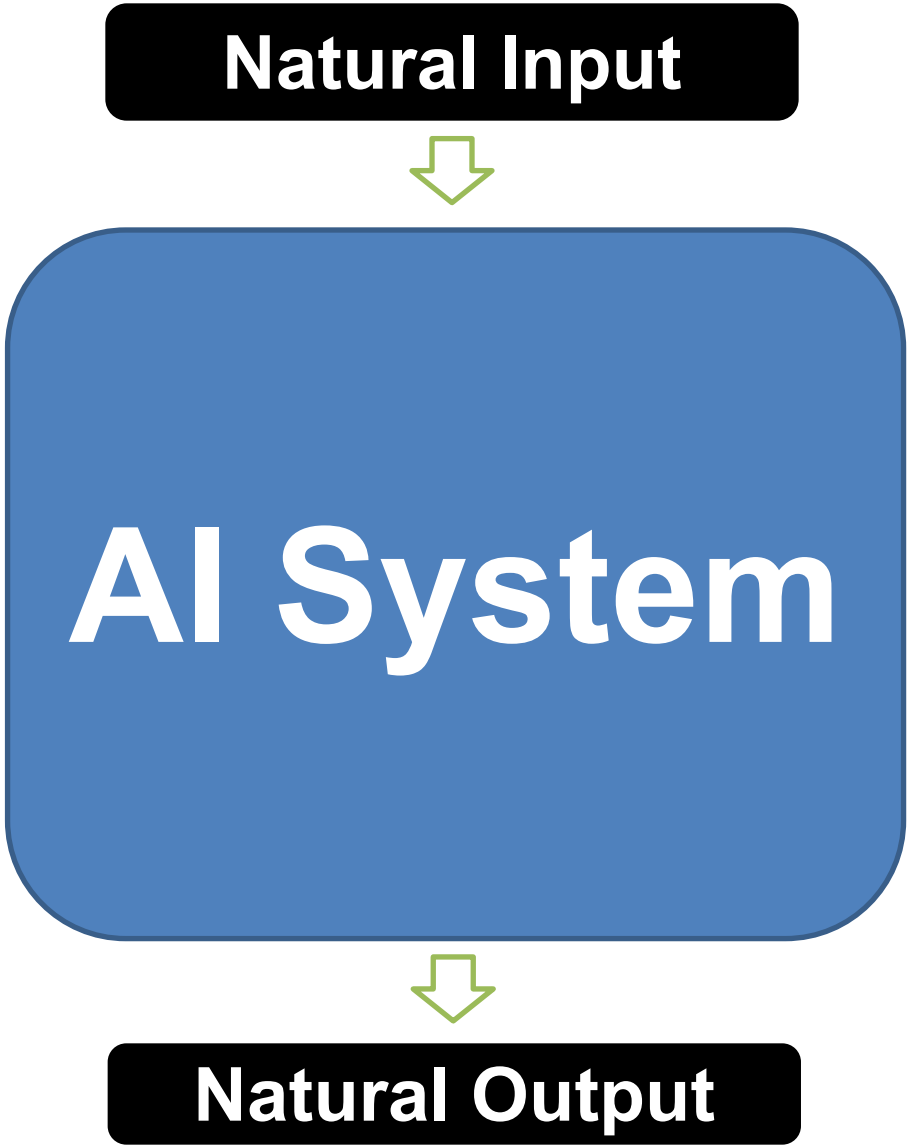
Tasks:

- **Question Answering and Reading Comprehension**
- **Textual Entailment**

1. Question Answering As Reasoning on Semantic Abstractions, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth, Submitted.
2. Learning What is Essential in Questions, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth, CoNLL, 2017
3. Relational Learning and Feature Extraction by Querying over Heterogeneous Information Networks, Parisa Kordjamshidi, Sameer Singh, D.K, StarAI, 2017
4. Better call Saul: Flexible Programming for Learning and Inference in NLP Parisa Kordjamshidi, D.K.,…, COLING, 2016.
5. Question Answering via Integer Programming over Semi-Structured Knowledge, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, …, IJCAI, 2016
6. EDISON: Feature Extraction for NLP, Simplified, Mark Sammons, Christos Christodoulopoulos, Parisa Kordjamshidi, D.K., … , LREC, 2016
7. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions, Peter Clark, Oren Etzioni, D.K., … AAAI, 2016.
8. Illinois-Profiler: Knowledge Schemas at Scale, Zhiye Fei, D.K., Haoruo Peng, Hao Wu and Dan Roth, Cognitum, 2015.
9. Solving Hard Co-reference Problems, Haoruo Peng, D.K. and Dan Roth, NAACL, 2015.
10. Flow of Semantics in Narratives, D.K, Chris J.C. Burges, Erin Renshaw, Andrzej Pastusiak, Tech Report, August 2014.

AI Goal:
Enabling machines to solve any problems, as good as human





"Yo …what's up?"

"Yo …not much!
Sup yourself?!"

Penn
University of Pennsylvania
Cognitive Computation Group

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."

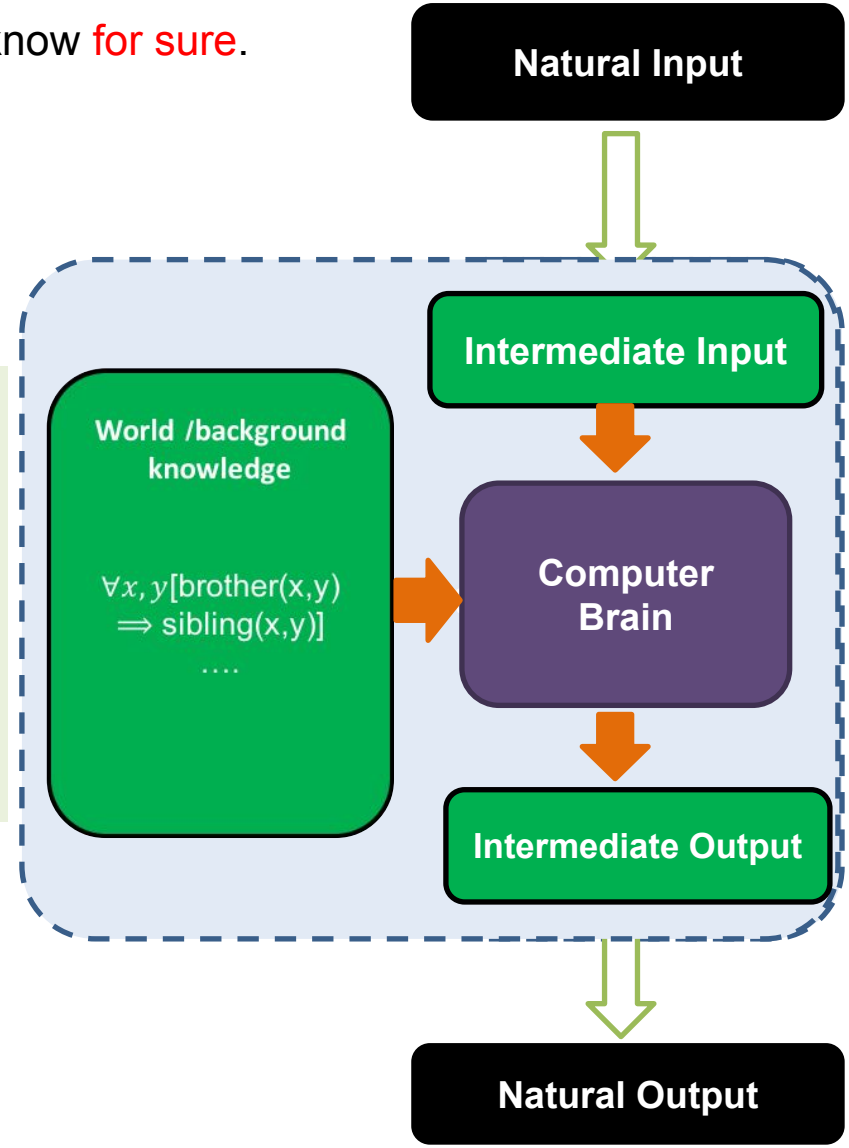
Reasoning: **Problem solving as Search!**

Still no unified model …

But there are certain things that we know for sure.

A "good" solution has to have:

- a Knowledge Representation (KR)
- knowledge.
- "easy" way of accessing the knowledge
- a decision making mechanism



Mary owns a canary named Paul. Does Paul have any ancestors who were alive in the year 1750?
(A) Definitely yes. (B) Definitely no. (C) There is no way to know.

ASK ME ANYTHING

"If you got a billion dollars to spend on a huge research project, what would you like to do?"

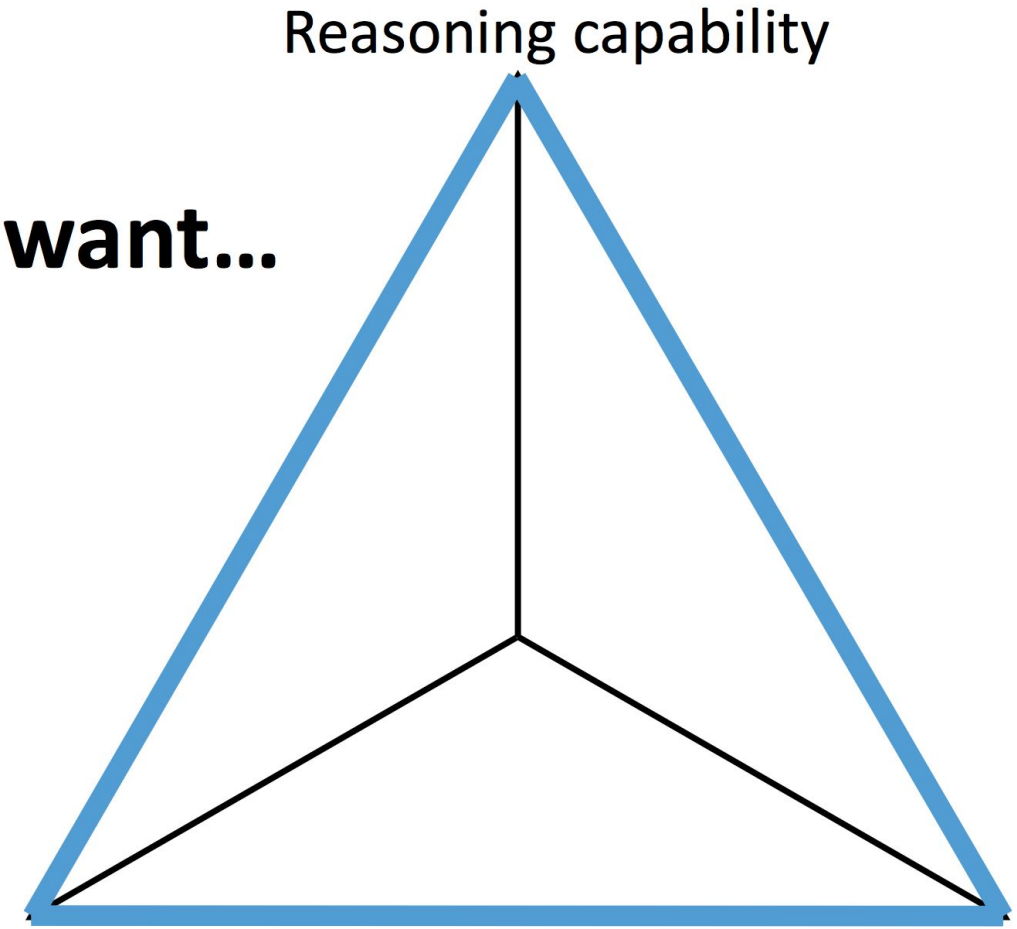
"I'd use the billion dollars to build a NASA-size program focusing on *natural language processing* (NLP), in all of its glory (semantics, pragmatics, etc)."

Michael Jordan
Professor of Computer Science
UC Berkeley

“Vague” line between non-reasoning QA and reasoning QA

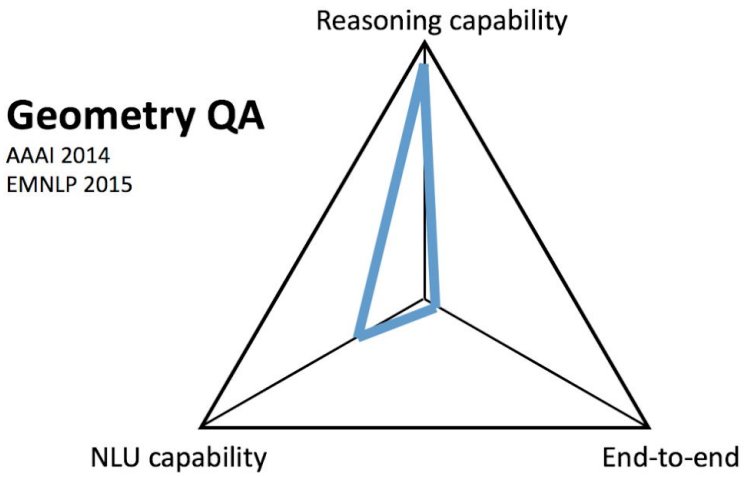
- Non-reasoning:
  - The required information is explicit in the context
  - The model often needs to handle lexical / syntactic variations
- Reasoning:
  - The required information may *not* be explicit in the context
  - Need to combine multiple facts to derive the answer
- There is no clear line between the two!

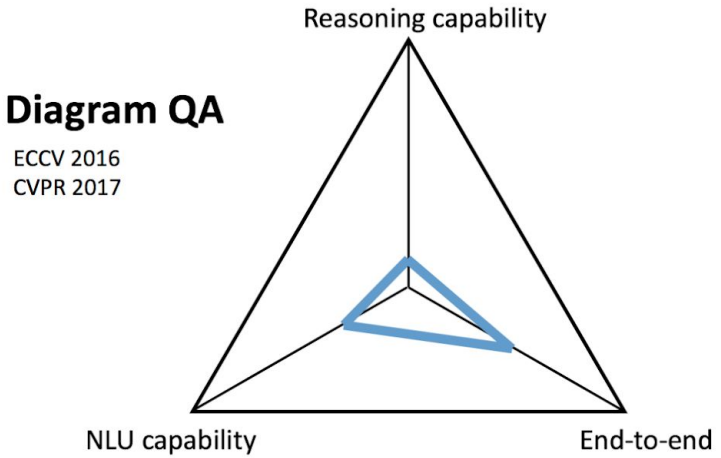
What we want…

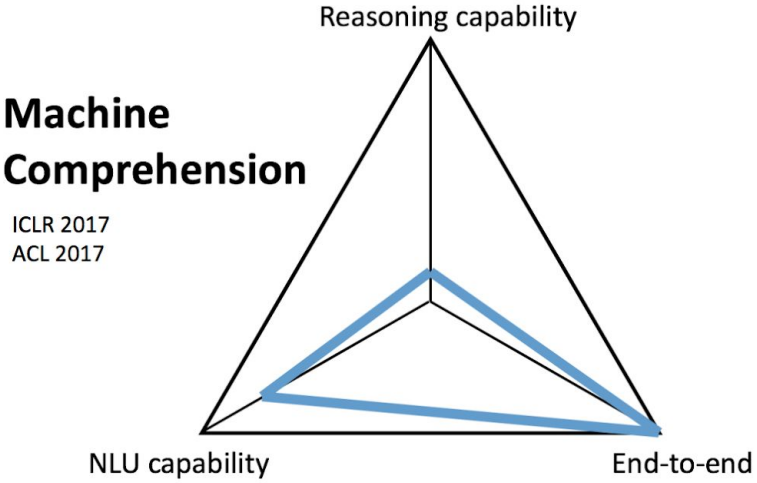


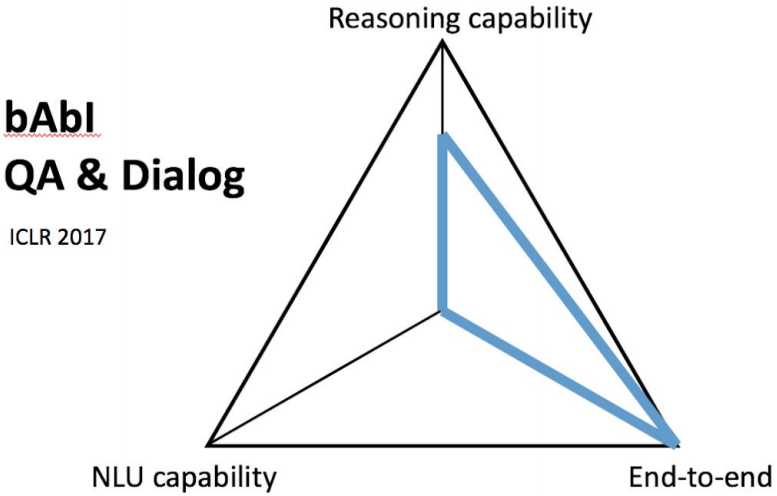
Three aspects of "reasoning system"

- **Natural language understanding**
  - How to retrieve relevant knowledge (formulas)?
  - Natural language has diverse surface forms (lexically, syntactically)
- **Reasoning**
  - Deriving new knowledge from the retrieved knowledge
- **End-to-end training**
  - Minimizing human efforts
  - Using only unstructured data









Cheeseburger stabbing

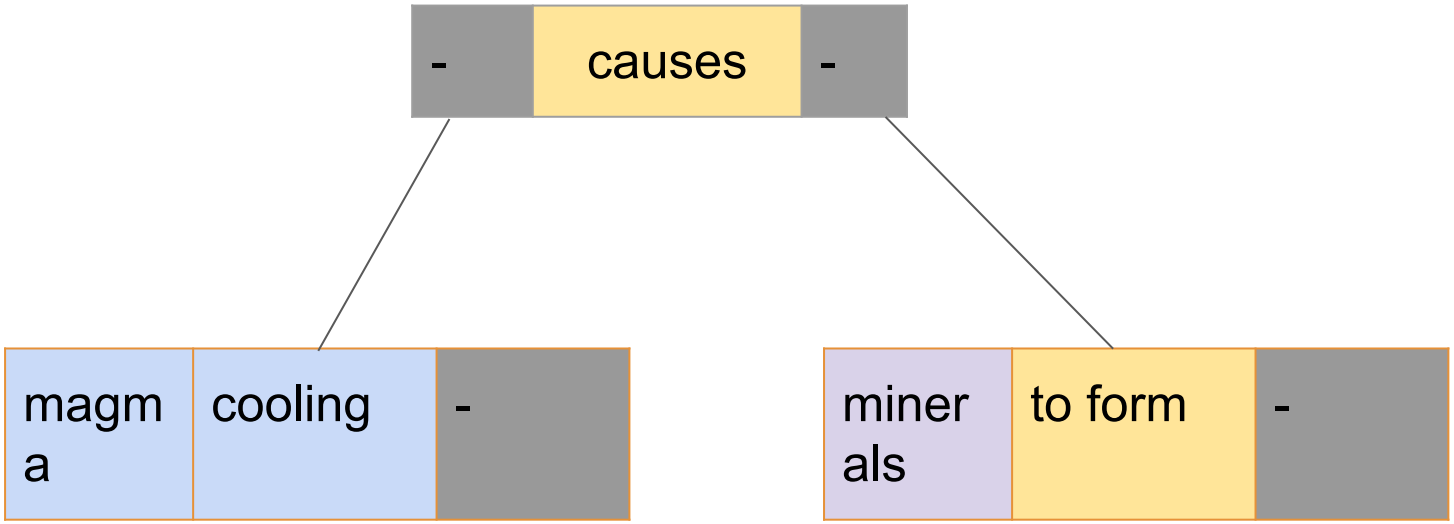
I can't decide if

- someone stabbed someone else over a cheeseburger
- someone stabbed someone else with a cheeseburger
- someone stabbed a cheeseburger
- a cheeseburger stabbed someone
- a cheeseburger stabbed another cheeseburger

Structure Matching

Challenge:
Second-order Paraphrase

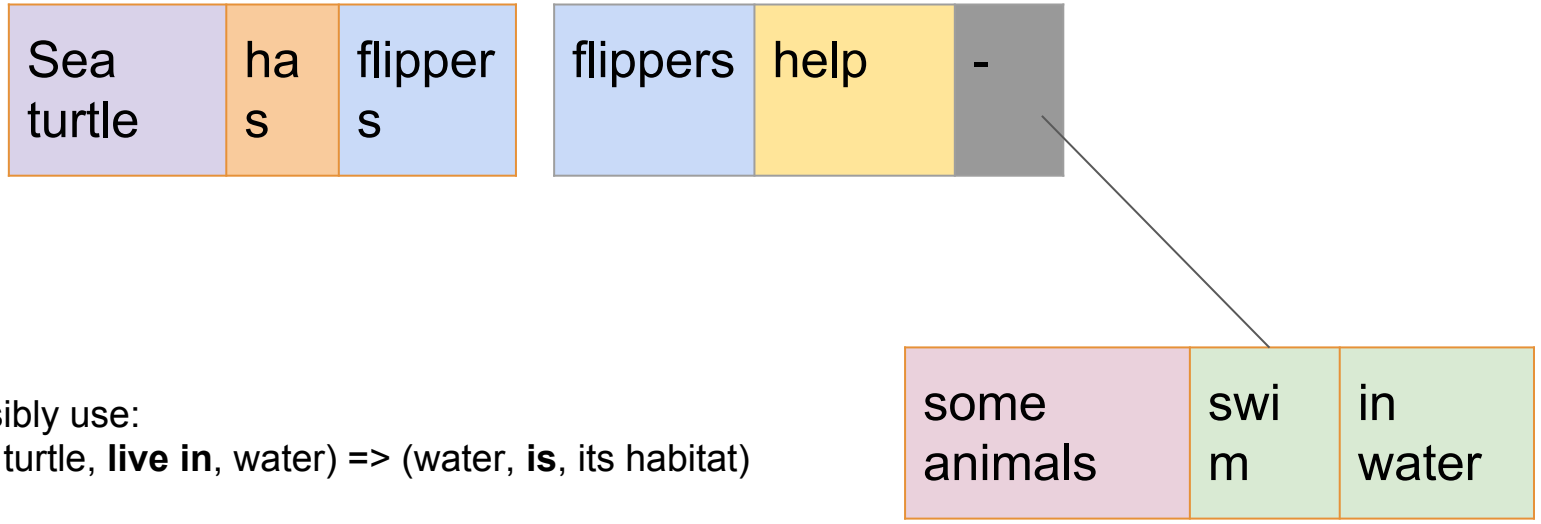
Minerals are formed by which process? (A) magma cooling (B) fault lines moving (C) metamorphosis (D) sedimentation



Simple Chaining

Challenge:
Recognizing chainable tuples

Which of the following animal features most helps the animal move around in its habitat? (A) A bird's sharp beak (B) A cow's tail (C) A sea turtle's flippers (D) A black bear's fur



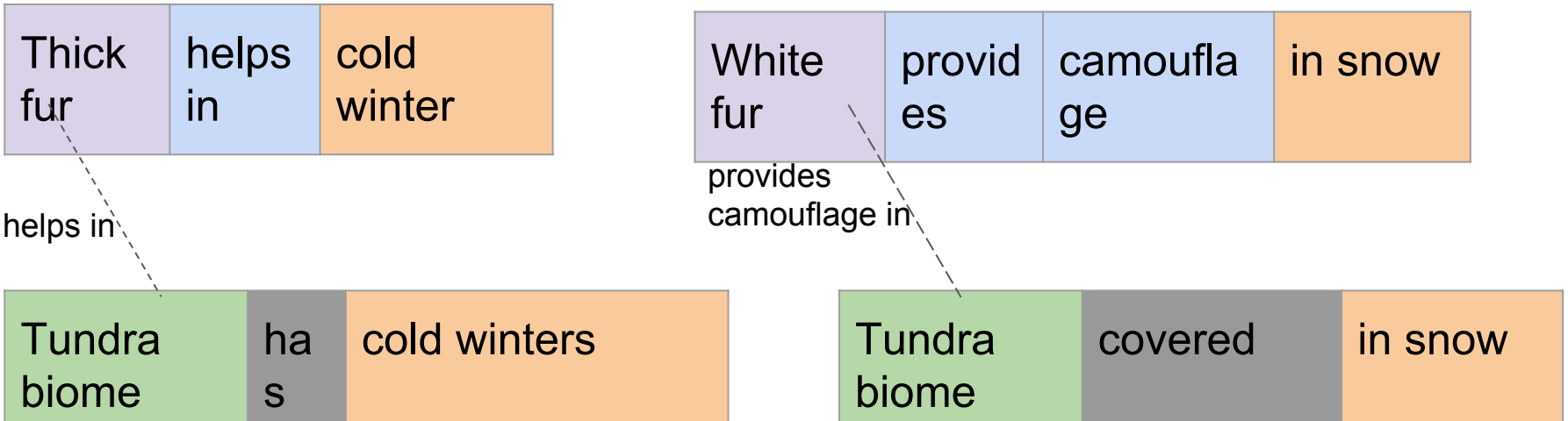
Possibly use:
(sea turtle, **live in**, water) => (water, **is**, its habitat)

Structure Chaining

Challenge:
Relations from chains

Thick white fur is an animal adaptation **most needed** for **the climate** in which biome? (A) deserts (B) taiga (C) deciduous forest (D) **tundra**

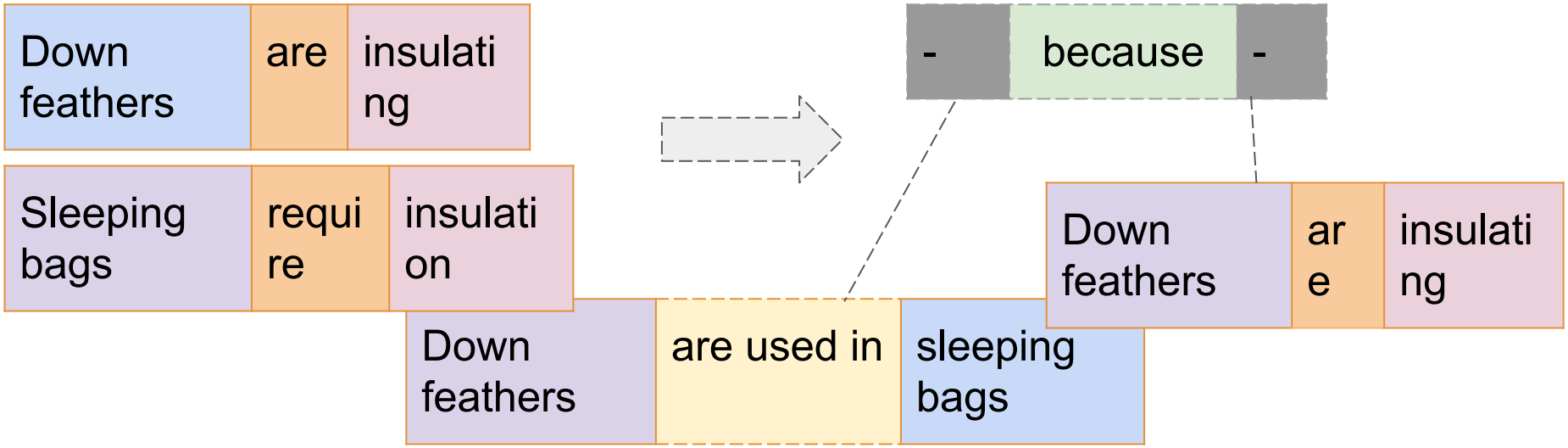
Type constrained rules:
(X, **helps in**, Y), (Z, **has**, Y) => (X, **helps in**, Z)
(X, **provides camouflage in**, Y), (Z, **covered in,** Y) =>
(X, **provides camouflage in**, Z)



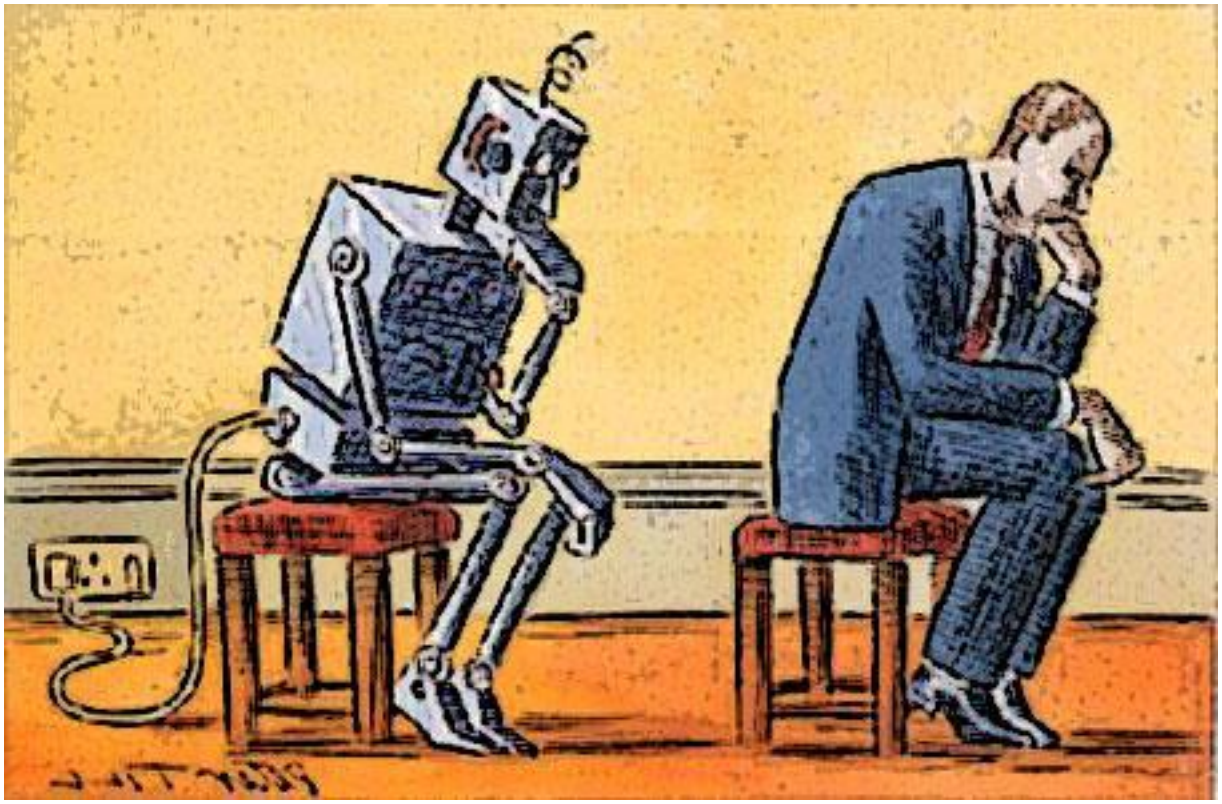
Structure Chaining

Challenge:
Second-order chain rules

Down feathers are used by many sleeping bag manufacturers because down feathers are (A) fire resistant. (B) comfortable padding. (C) good insulators. (D) water resistant.



Machine Intelligence



"It would someday be possible for a sufficiently advanced computer to think and to have some form of consciousness"

-- Computing Machinery and Intelligence, Mind 1950.

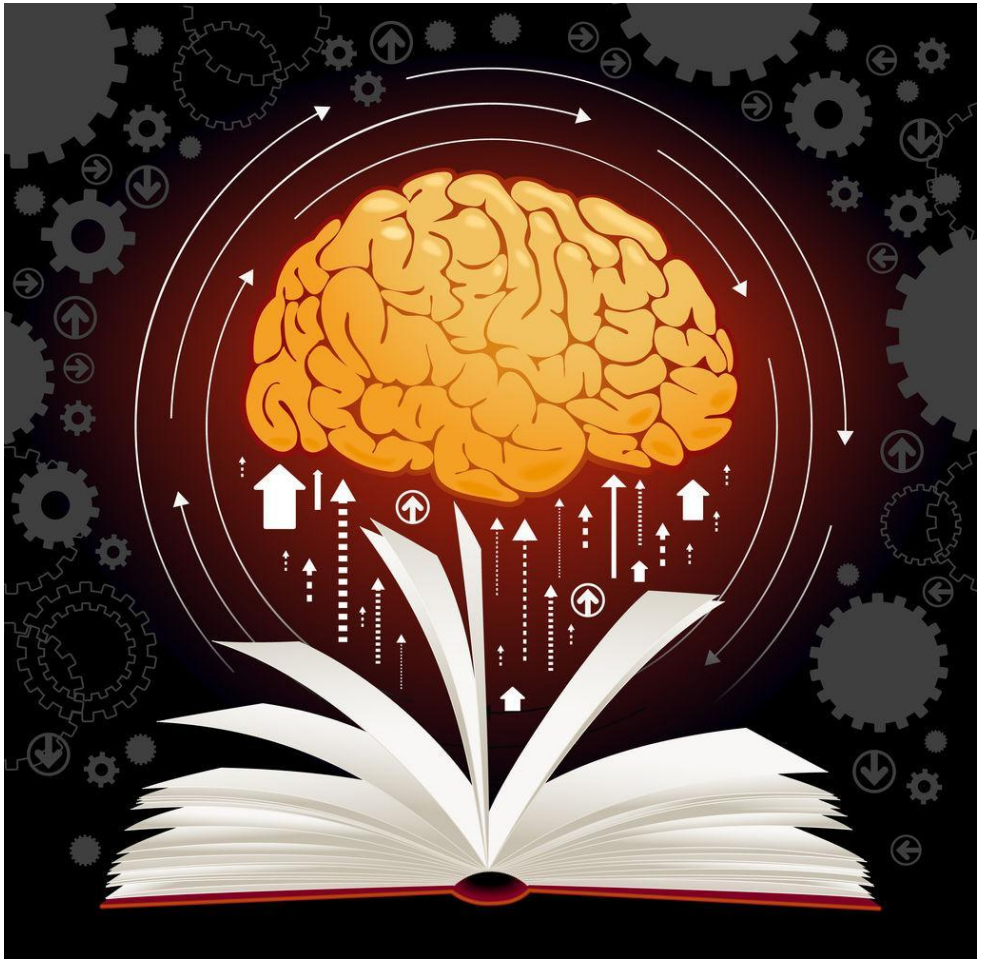
How do we measure progress? What tasks should drive the field?

Turing Test

Standardized Tests as drivers for AI ?

-- [Levesqeue 2010, Clark 2014]

Standardized Tests



Why Standardized Tests

- *Easily accessible*
- *Easily measurable*
- *Do not cover all aspects of intelligence at once*