Reasoning-driven Question Answering

Daniel Khashabi

Stanford NLP Seminar
June 28, 2018

@DanielKhashabi

Programs with Commonsense

[John McCarthy, 1959]

Formalize world in **logical** form!

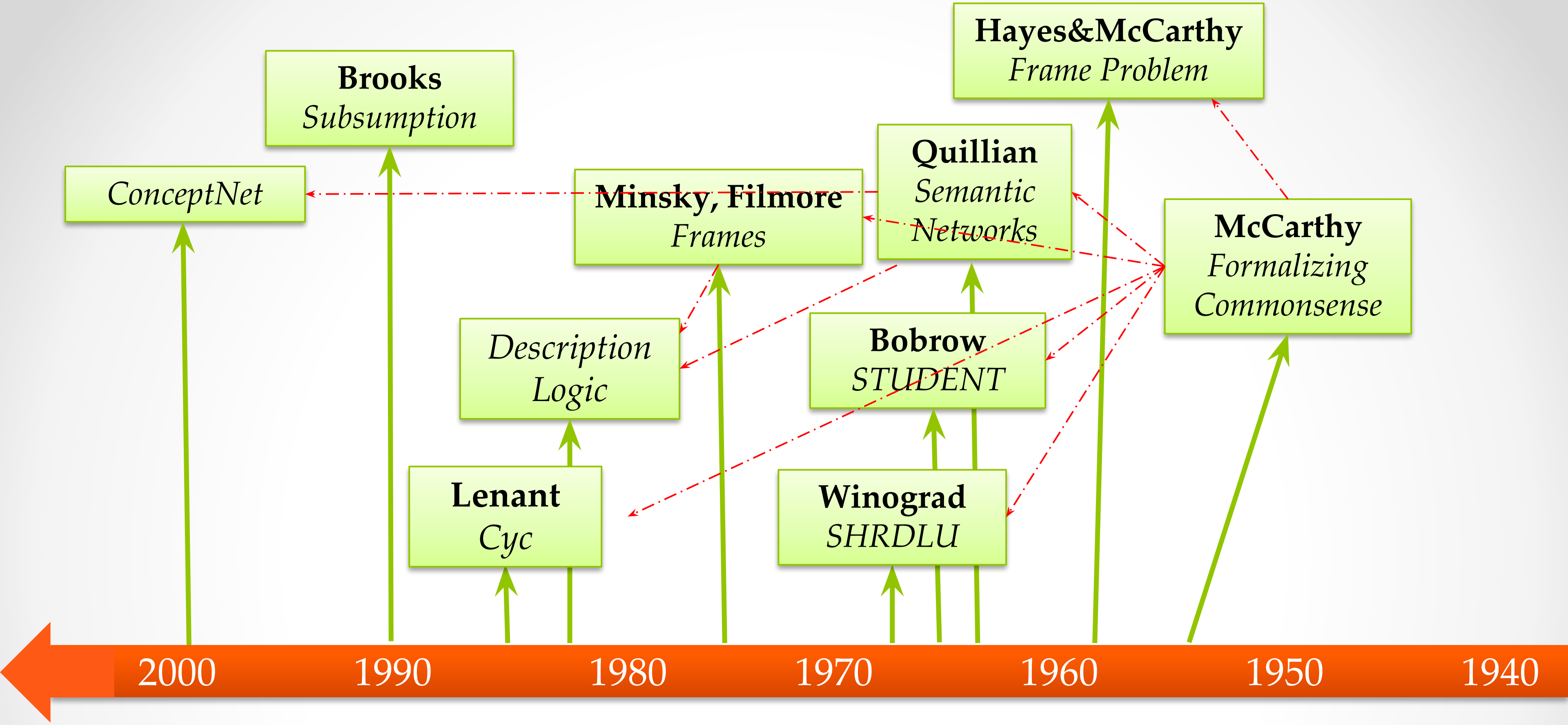
Example:
"My desk is at home" → at(I, desk)
"Desk is at home" → at(desk, home)

Hypothesis: Commonsense knowledge can be formalized with logic.

Do **reasoning** on formal premises!

Example Contd.:
$\forall x \forall y \, \forall z$ at(x,y), at(y,z)→ at(x, z)
∴ at(I, home)

Hypothesis: Commonsense problems are solved by logical reasoning



They were right that, once you understand language, you can do reasoning;

but they underestimated the difficulty of NLU.

Variability and Ambiguity

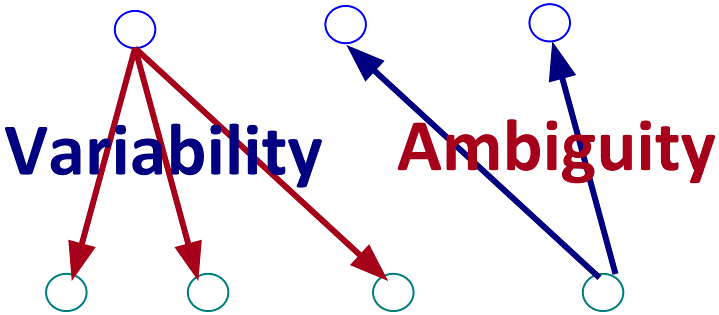
- The difficulty of mapping from nature (including natural language) to symbols

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

[The Symbol Grounding Problem, S. Harnad, 1990]

“Chicago”

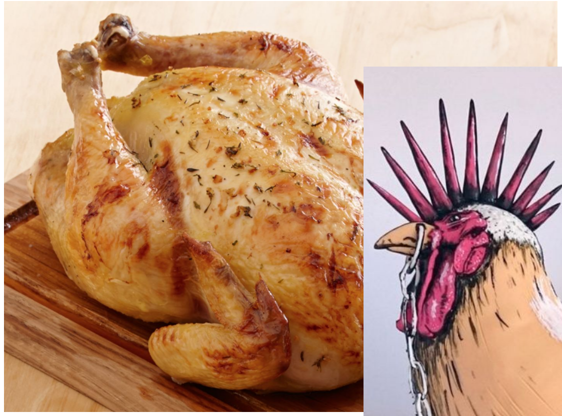
Meaning



Language

Structural Ambiguity

Chickens are ready



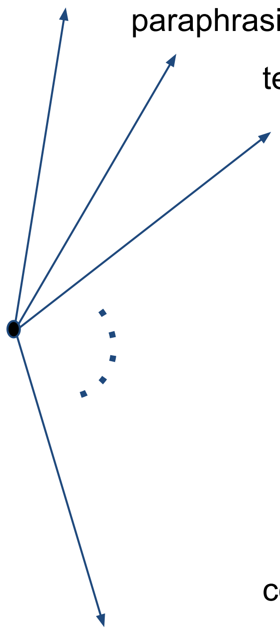
+ to eat

The many faces of reasoning

- Reasoning is often studied in a very narrow sense.

Reasoning has many (infinite?) forms.

- Examples typically span multiple reasoning aspects.



quantitative reasoning
paraphrasing
temporal
deductive
inductive
causal (cause to effect)
causal (effect to cause)
abductive
analogy
exemplar (learn. by ex.s)
conditional
non-monotonic
coref
....

The many faces of reasoning

Abductive reasoning

Incomplete Observations → Best conclusion (maybe true)

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

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

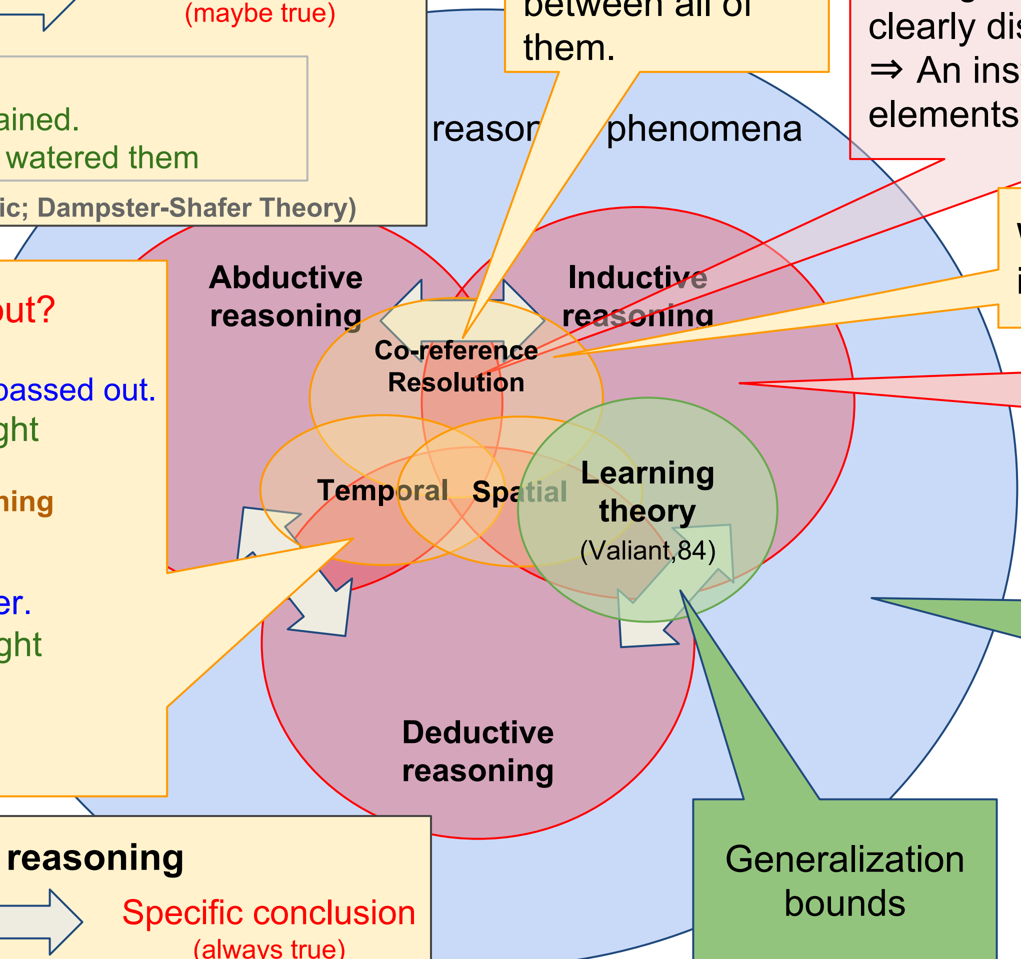
Q: When did Jack pass out?

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

⇒ **Abduction: (probably) morning**

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

⇒ **Deduction: night**



There is overlap between all of them.

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

What a linguist would interpret “reasoning”

What a logician would interpret as “reasoning”

Very little understanding

Generalization bounds

Deductive reasoning

General Rule → Specific conclusion (always true)

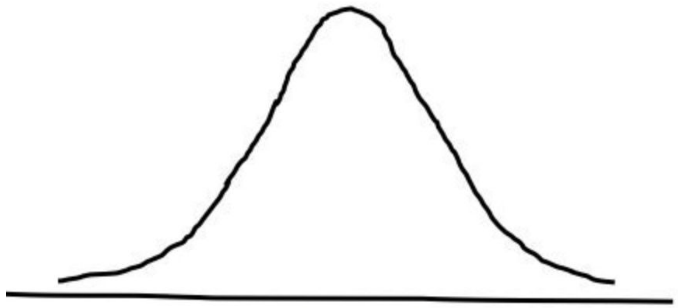
When it rains, objects get wet.
It rained.
- The grass must be wet.

(modus ponens; modus tollens)

PAC learning and generalization

[Valiant, 1984]

Hypothesis class H (e.g. a neural net)



Latent Distribution

train

test

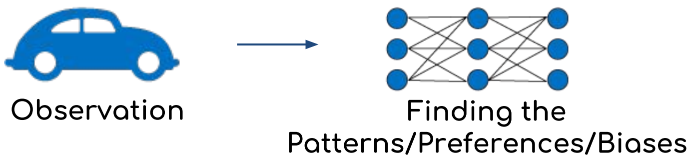
A hidden consistent concept

$h \in H$

$$\underbrace{\mathbf{P}\overbrace{\left(\mathrm{error}_{\mathrm{test}}(h) \leq \epsilon\right)}^{\textbf{Approximately}} \leq 1-\delta}_{\textbf{Probably}}$$

Not everything is (easily) inductively learnable

The dominant approach to "learning" is by "many observations".



- Close to how induction works:

Observation → Pattern → Hypothesis

- Not a good induction.

- Many problems that might not be easy to be solved with induction:
  - Math word problems
  - Fiction story understanding

In fact you can't even create big enough training set for them.

John had $\boxed{6}$ books; he wanted to give them to two of his friends. How many will each one get?

- Sensitive to deviations from the dominant bias (aka *adversarial examples*)

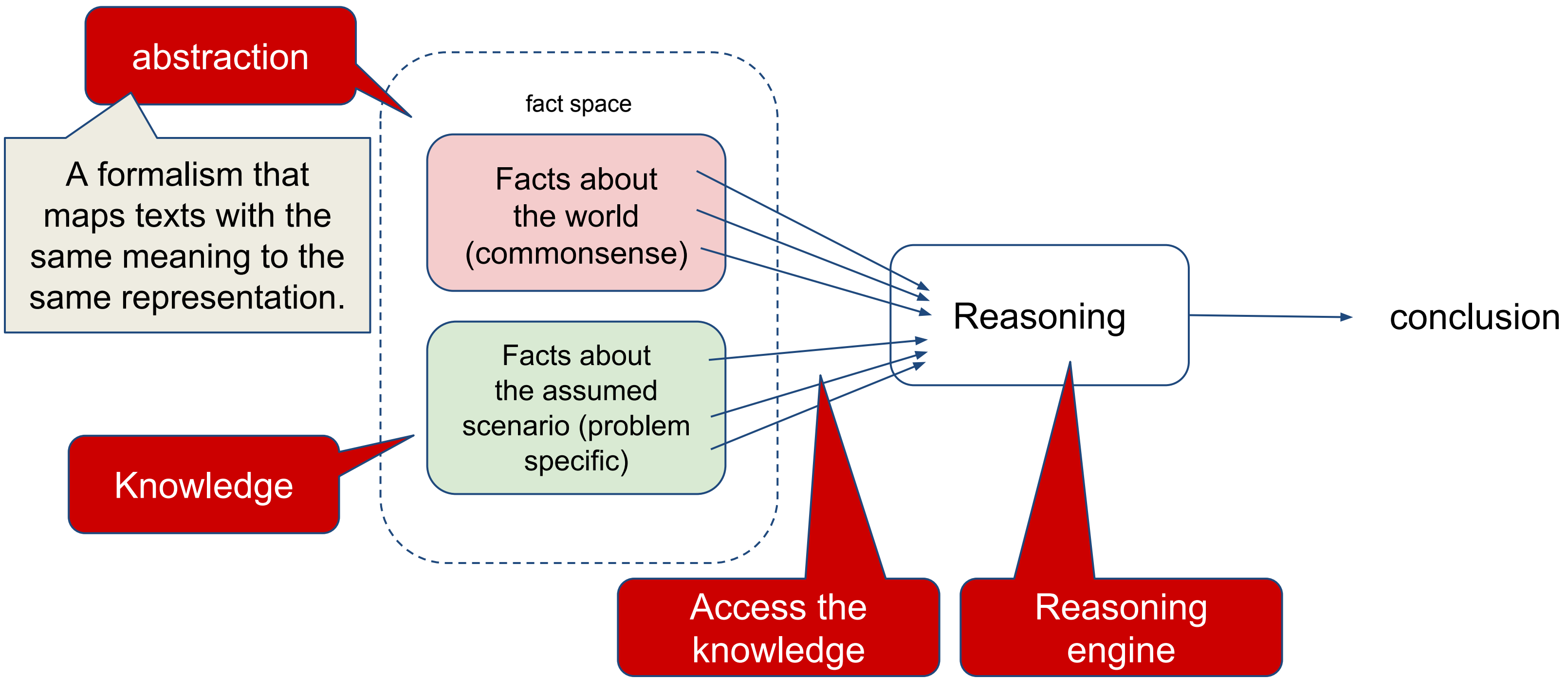
A turkey, fed every morning without fail,who following the laws of induction concludes this will continue, but then his throat is cut on Thanksgiving Day.

--Bertrand Russell

Talk statement

- Question answering is a natural language understanding problem.
- Automating natural language understanding requires reasoning.
- Effective reasoning requires a wide spectrum of inter-dependent abilities working together coherently.

The big picture



Roadmap

- ❖ Motivation & Background

- ❖ **Reasoning-Driven Question Answering**

  System Design Aspect

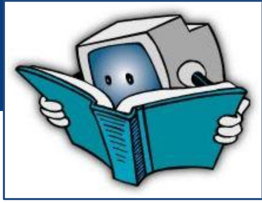
  ➢ Global Reasoning Over Semantic Abstractions (IJCAI’16, AAAI’18)

  Evaluation Aspect

  ➢ A Challenge Set for Reasoning Over Multiple Sentences (NAACL’18)

- ❖ **Concluding Remarks**

Exams



Standardized science exams (Clark et al, 2015):

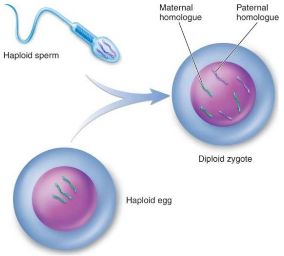
- Simple language; kids can solve them well, but they need to have the ability use the knowledge and abstract over it.

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.



Linguistic 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 in 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 a NLU problem

A single abstraction is not enough

High-level view

Question Answering

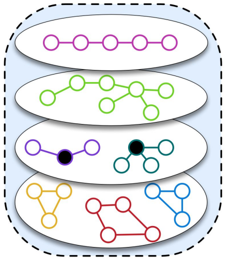
as Global Reasoning

over Semantic Abstractions

Question Answering

Semantic Abstractions

Global Reasoning



Collections of semantic graphs

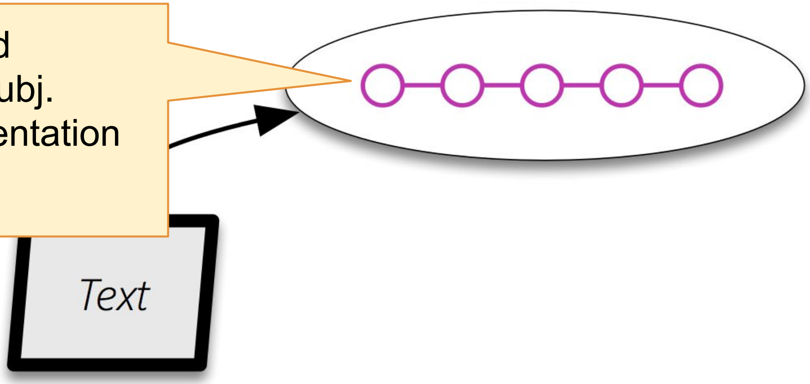
Create a unified representation of families of graphs

- 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

...



SemanticGraph 1
(Sequence)

e.g named-entities

e.g dependency parse

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

e.g co-reference

5

e.g tables

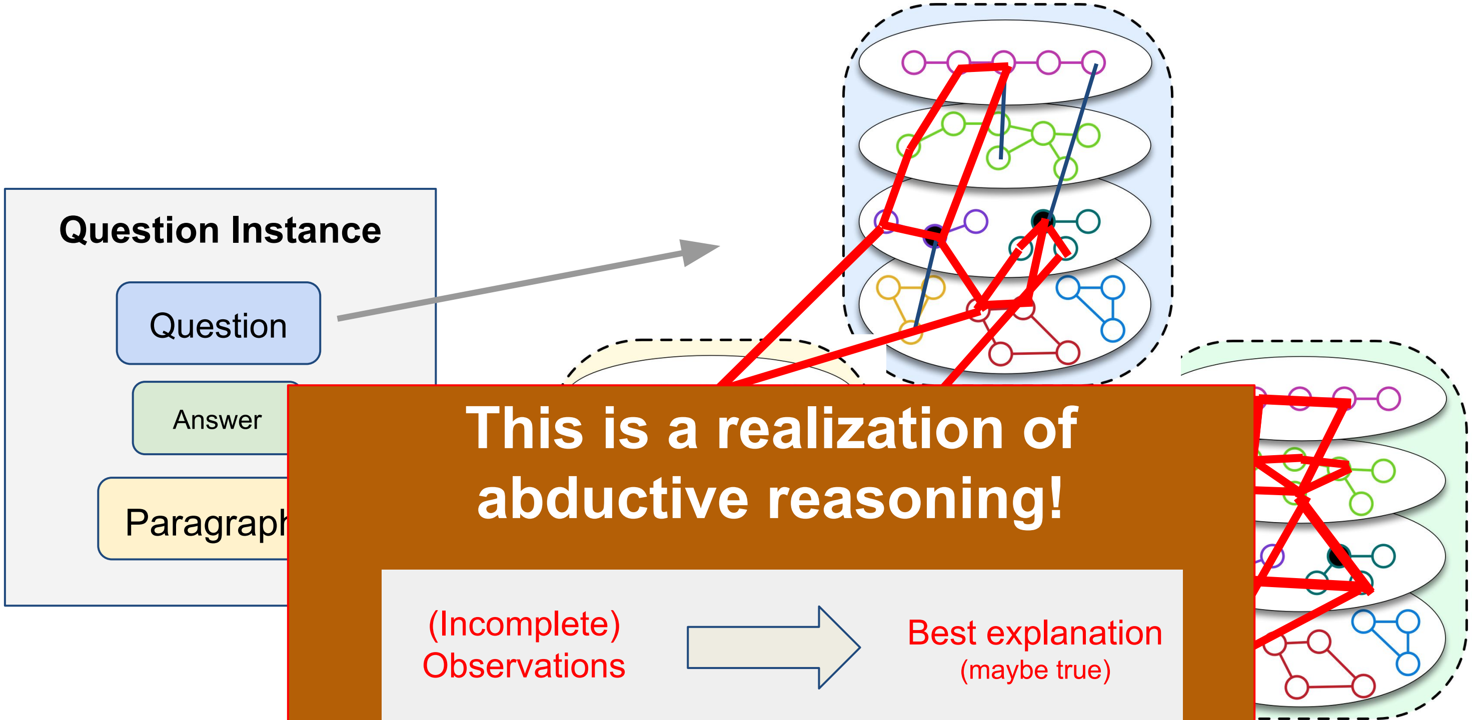
TableILP: IJCAI'16

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

Edges reflect similarity / entailment

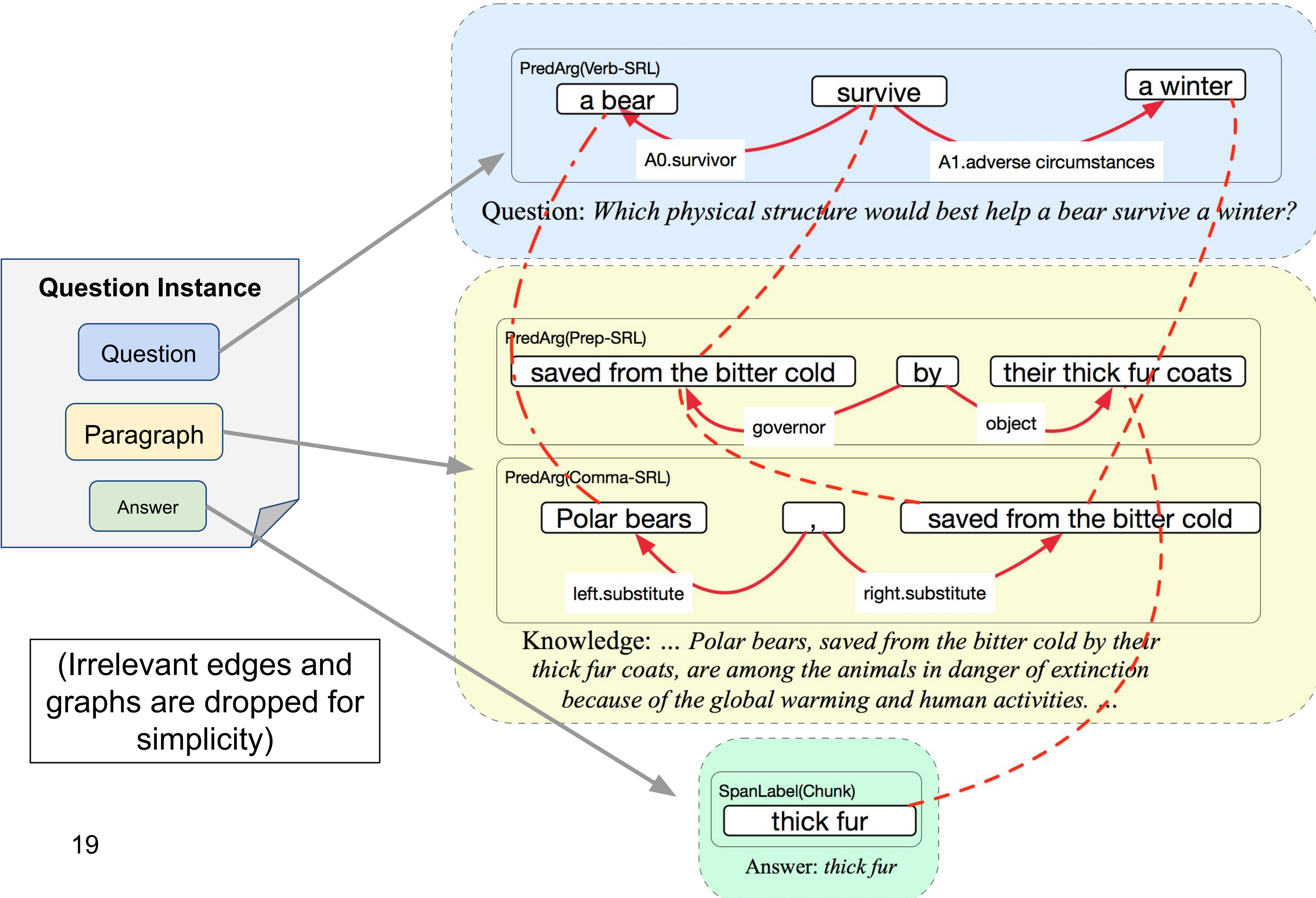


This is a realization of abductive reasoning!

(Incomplete) Observations → Best explanation (maybe true)

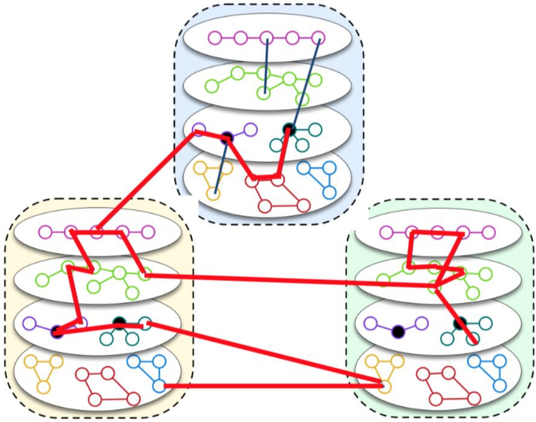
QA Reasoning formulated as finding "best" explanation – subgraph connecting Q to A via P

Example subgraph



(Irrelevant edges and graphs are dropped for simplicity)

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 a valid reasoning, what's preferred

- **Preferences** e.g.
  - Use sentences nearby
  - If using a pred-arg graph, give priority to the subject

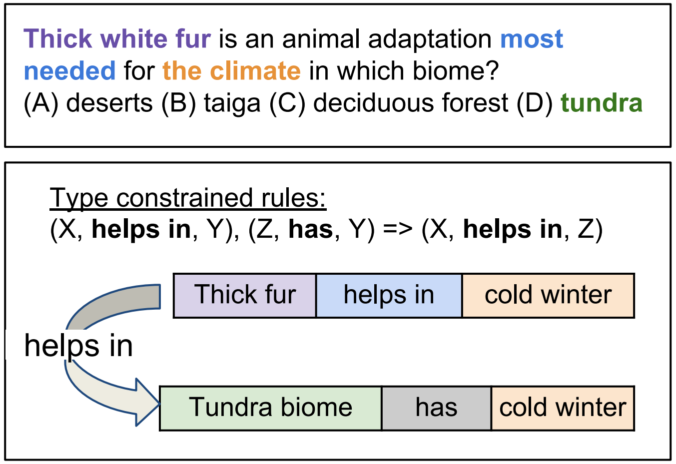
Formulate as Integer Linear Program (ILP) optimization

- Solution points to the best supported answer

Evaluation: notable baselines

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

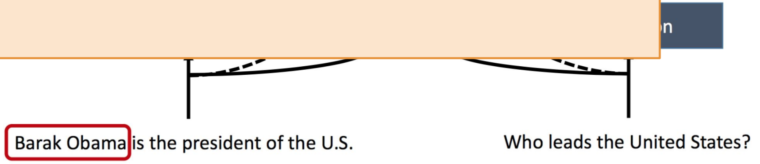
- TupleINF **(Khot et al, ACL'17)**
  - Inference over **independent rows**
  - **Auto-generated short triples**
  - And type-constrained rules



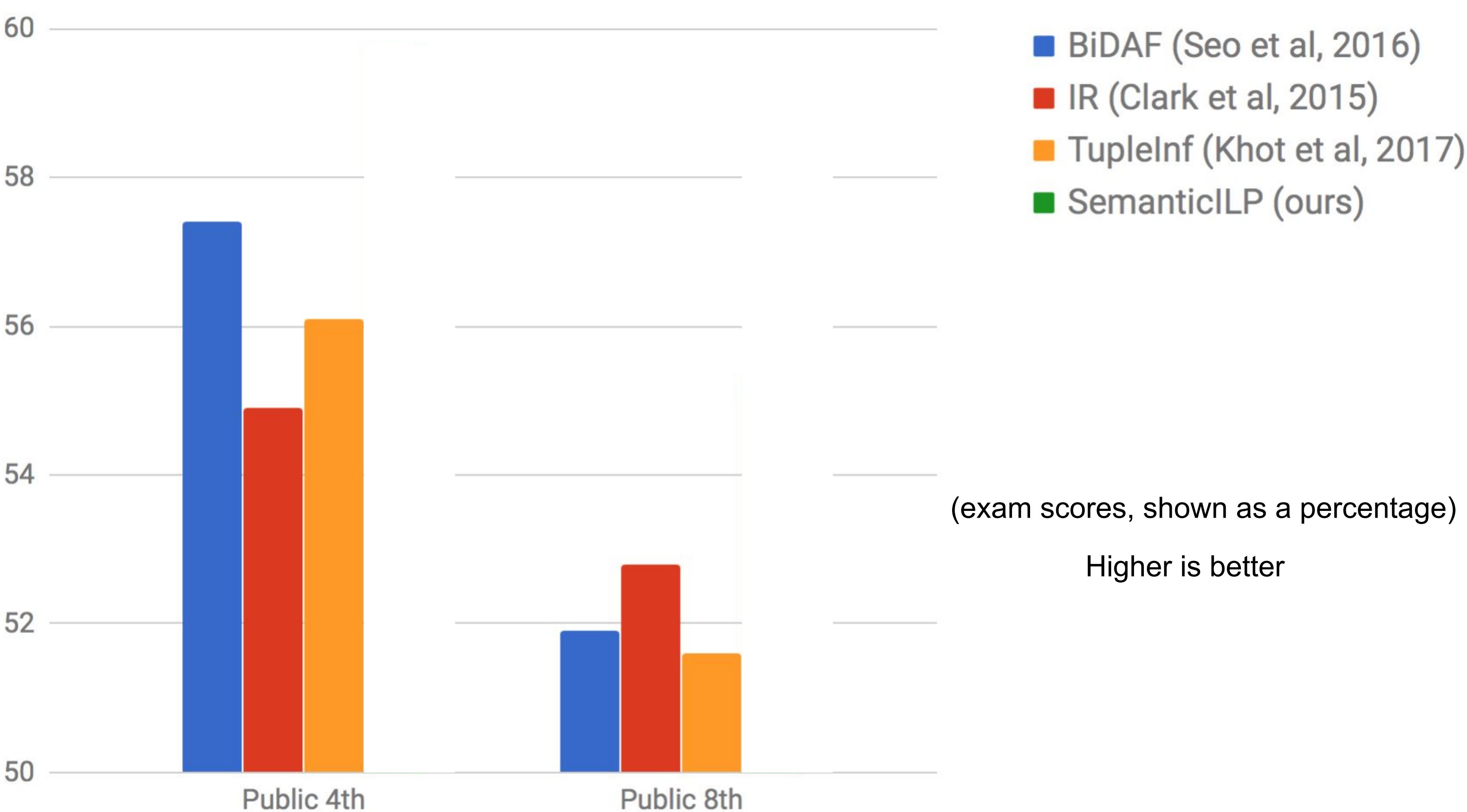
- BiD

We compare with the best baseline on each domain.

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



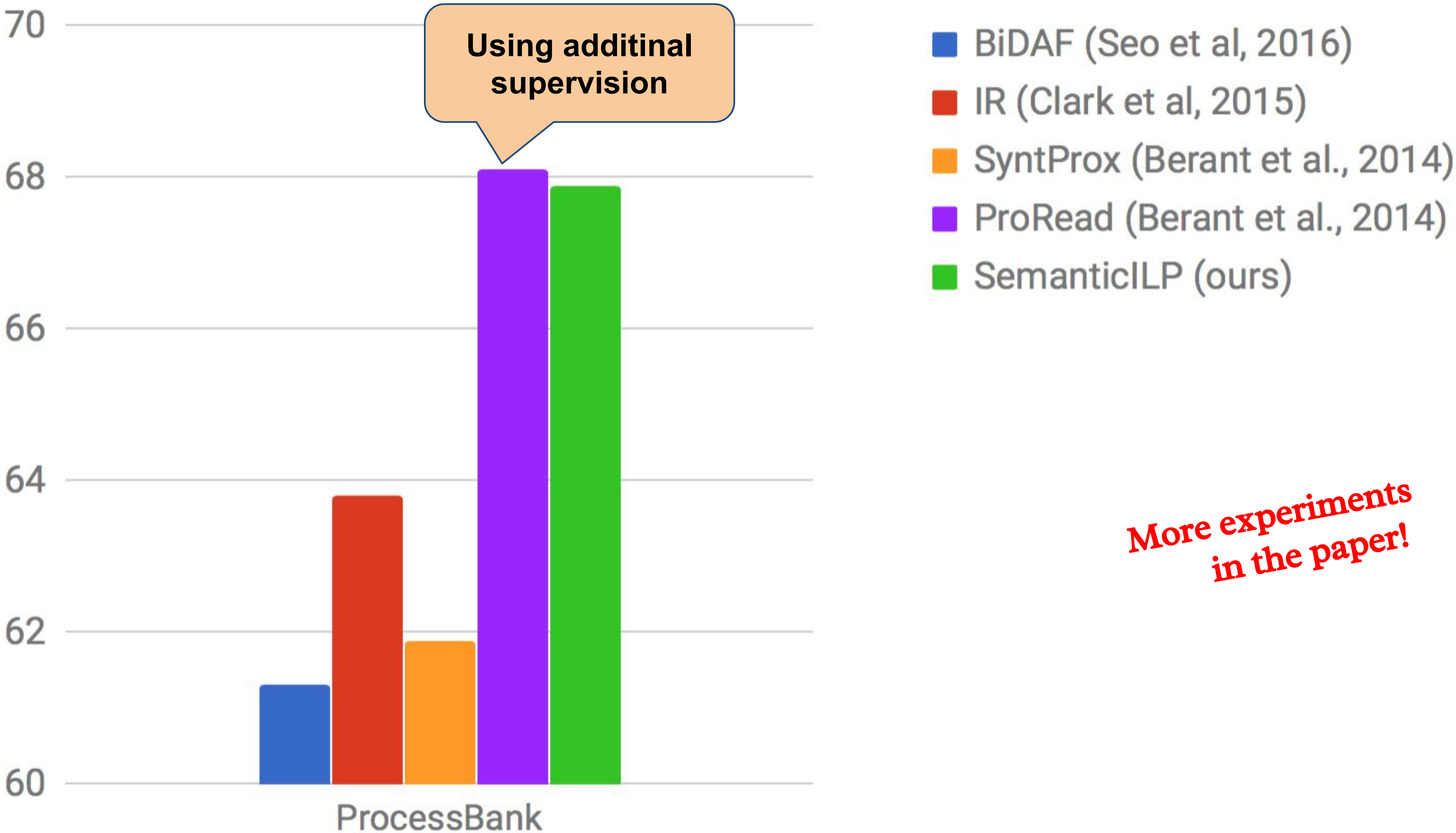
Results #1: Science Questions



(exam scores, shown as a percentage)

Higher is better

Results #2: Biology Questions



More experiments in the paper!

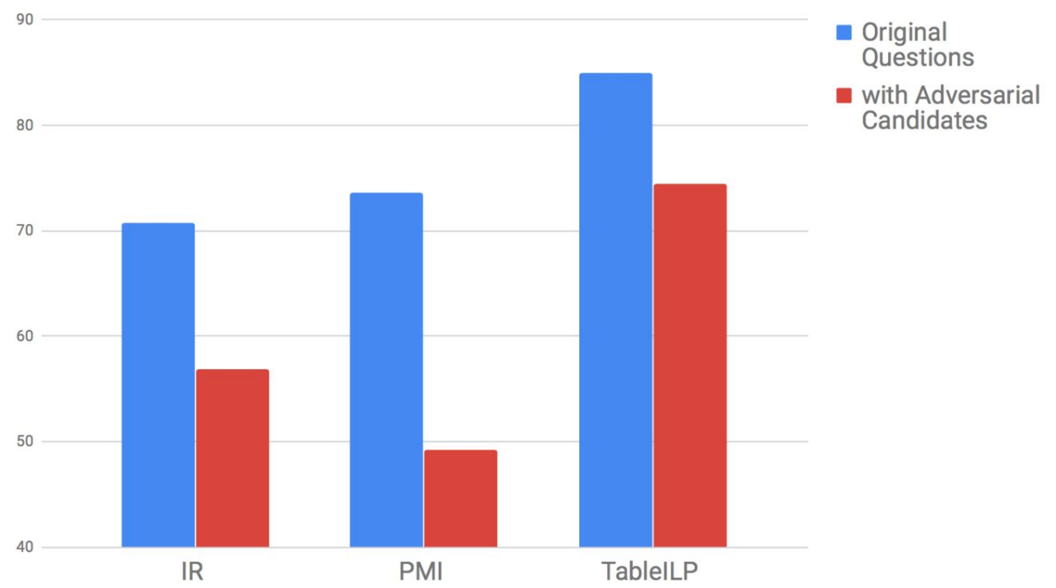
One single system tested on different datasets.

Assessing Brittleness: 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*

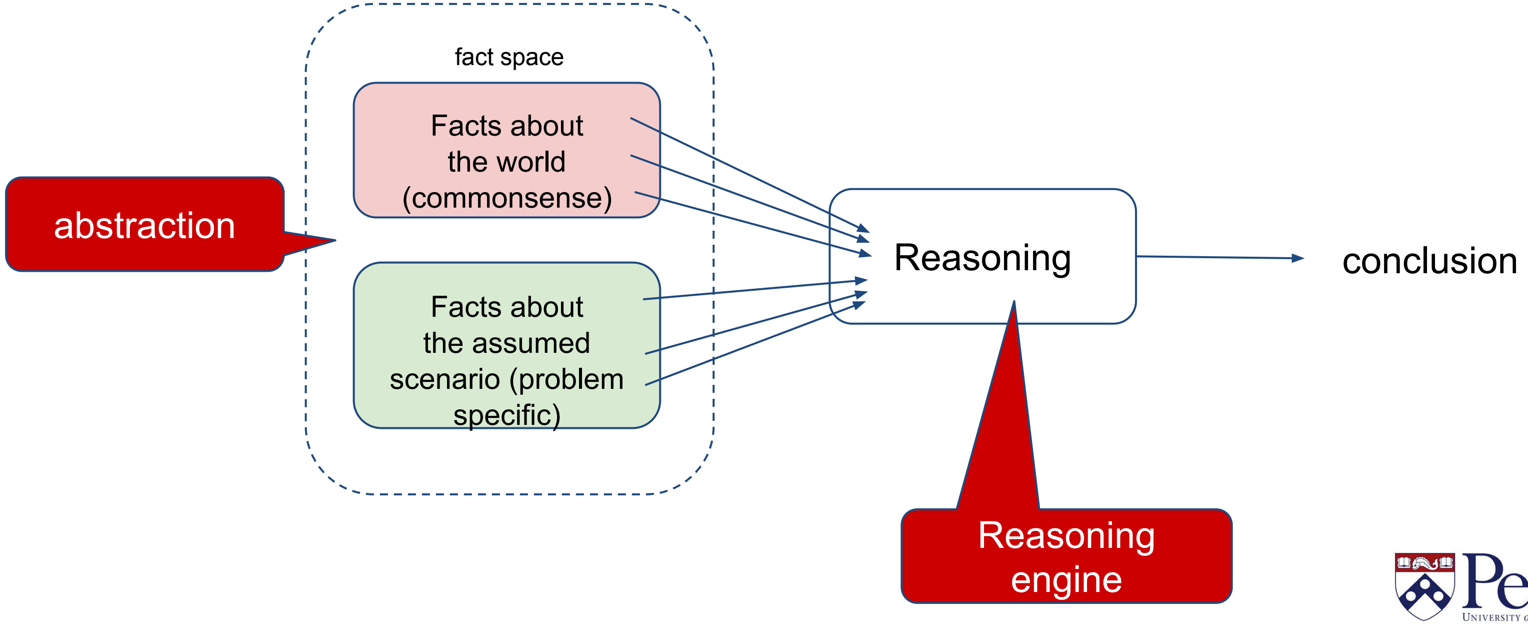


[IJCAI'16]

[Jia&Liang,EMNLP'17]

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**
- ❖ **Reasoning-Driven Question Answering**

System Design Aspect

➢ Global Reasoning Over Semantic Abstractions (IJCAI'16, AAAI'18)

Evaluation Aspect

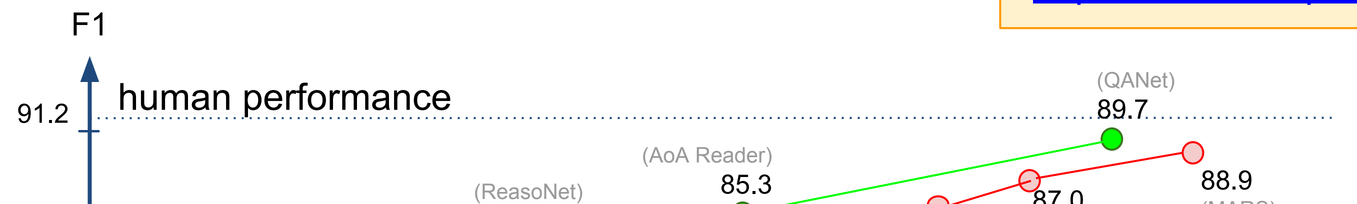
➢ A Challenge Set for Reasoning Over Multiple Sentences (NAACL'18)

- ❖ **Concluding Remarks**

Rapid progress on SQuAD

Stanford Question Answering Dataset

https://stanford-qa.com



OBSERVER

TECHNOLOGY | ECONOMY | STARTUPS | PERSO

Alibaba, Microsoft AI Programs Beat Humans on Reading Comprehension Test

By John Bonazzo • 01/16/18 11:47am

Will the artificially intelligent robot from Ex Machina become a reality? Steve Troughton/Flickr Creative Commons

Artificial intelligence has improved by leaps and bounds in recent years, able to help with household chores and judge beauty contests. And now AI programs

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

- **The resulting systems are brittle**

[IJCAI’16]

[Jia&Liang,EMNLP’17]

Penn
University of Pennsylvania

Overfitting to the dataset generation process

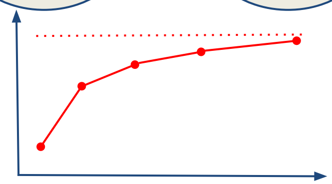
Dataset generation process

distribution of the underlying dataset generation process

the actual task distribution

test

train



Successfully managed to learned the distribution, but …

The goal is to learn “tasks”, not an approximate distribution.

Annotator objective:
maximizing profit, while following the task guidelines

Inducing “reasoning” in a dataset

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 restricted

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

“Multi-sentence” hypothesis:

Questions that require multiple sentences tend to be “hard”.

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

MultiRC: Reasoning over multiple sentences.

A reading comprehension challenge set with questions that require 'reasoning' over more than one sentence 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

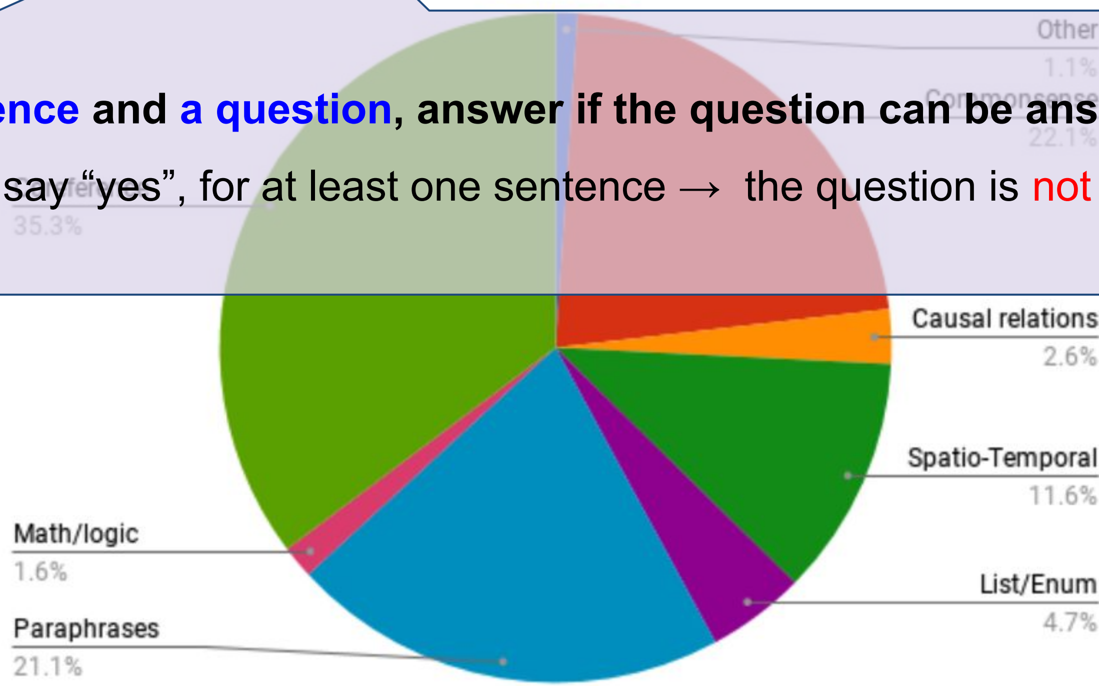
"know what don't know" [Rajpurkar,Jia& Liang, ACL'18]

MultiRC: Question generation pipeline

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

Generate multi-sentence questions → **Verify multi-senten ce-ness** → **Generate candidate correct / incorrect answers** → **Dataset quality verification**

Phenoma breakdown

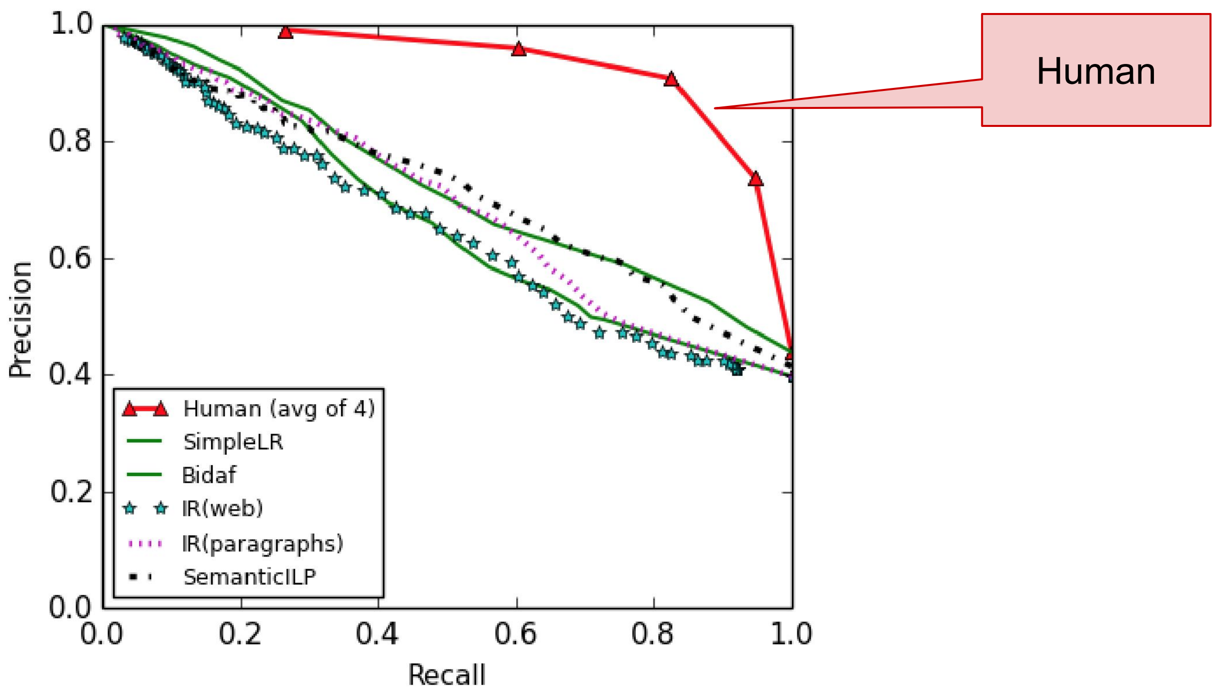


Given a sentence and a question, answer if the question can be answered

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

Baseline performances

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



Reusability of test set

In principle the test set should be used only once.

Leaderboard participants are allowed to repeatedly evaluate their submissions

They may begin to overfit to the holdout data, over time.

- Alternatives to "best submission of each team" strategy
  - Adaptive strategy to approximate unbiased estimate of the true performance

[Dwork et al, 2015] [Blum&Hardt, 2015]

Our solution:
Every few months we will include a new unseen additional evaluation data

Release Tag	Release Date	Released?
R1	Spring, 2018	✔
R2	Winter, 2019	×
R3	Summer, 2019	×
R4	Fall, 2019	×

cogcomp.org/multirc

Homepage Releases and downloads Explore the data CodaLab Evaluation Script

MultiRC

Reading Comprehension over Multiple Sentences

Introduction

MultiRC (Multi-Sentence Reading Comprehension) is a dataset of short paragraphs and multi-sentence questions that can be answered from the content of the paragraph.
We have designed the dataset with three key challenges in mind:

- The number of correct answer-options for each question is not pre-specified. This removes the over-reliance of current approaches on answer-options and forces them to decide on the correctness of each candidate answer independently of others. In other words, unlike previous work, the task here is not to simply identify the best answer-option, but to evaluate the correctness of each answer-option individually.
- The correct answer(s) is not required to be a span in the text.
- The paragraphs in our dataset have diverse provenance by being extracted from 7 different domains such as news, fiction, historical text etc., and hence are expected to be more diverse in their contents as compared to single-domain datasets.

The goal of this dataset is to encourage the research community to explore approaches that can do more than sophisticated lexical-level matching.

Leaderboard

Here we show a summary of the best results on our dataset:

System	Paper	Dev		Test(R1)	
		F1m	F1a	F1m	F1a
Human (avg of 4)	(Khashabi et al, 2018)	86.40	83.80	84.32	81.82
Logistic Regression	(Khashabi et al, 2018)	66.08	63.77	66.68	63.46
Information Retrieval	(Khashabi et al, 2018)	64.25	60.04	54.83	53.94
Random baseline	(Khashabi et al, 2018)	46.12	46.74	47.11	47.57

To see our evaluation script and a few baseline scores take a look at this repository. For instructions on how to evaluate your system,

Summary

- We need reading comprehension playground which requires deeper "reasoning"
- An approach proposed here: enforcing dependence on multiple sentences.

Beyond this work:

- Different communities evaluate on different datasets
- Let's evaluate on multiple datasets
- A dataset being small is not an excuse for not using it.

Roadmap

- ❖ **Motivation & Background**
- ❖ **Reasoning-Driven Question Answering**

System Design Aspect

➢ Global Reasoning Over Semantic Abstractions (IJCAI'16, AAAI'18)

Evaluation Aspect

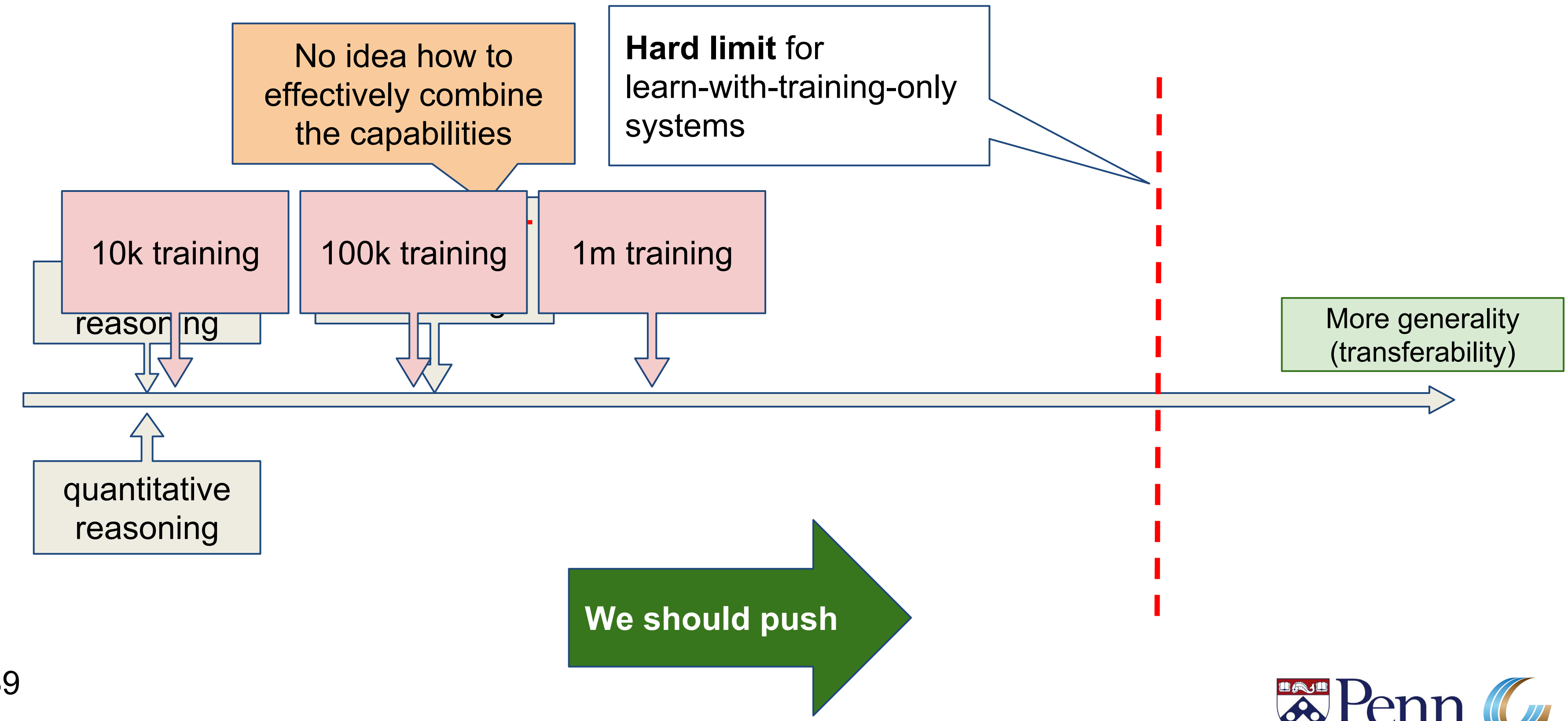
➢ A Challenge Set for Reasoning Over Multiple Sentences (NAACL'18)

- ❖ **Concluding Remarks**

Recap

- Studying “reasoning” is a crucial element towards solving QA.
- We studied a few aspects of reasoning:
  - **System design:**
    - An abductive model, on top of *semantically-informed* representation.
  - **Evaluation:**
    - A playground that will force us to address reasoning when we study QA.
- What’s missing:
  - ?

- For a “good” QA there is no notion of domain or dataset.
- Reasoning shouldn’t be defined too narrowly
- Language understanding should not be equated with training on datasets.



Acknowledgement

Dan Roth (UPenn)

Snigdha Chaturvedi
(UCSC)

Tushar Khot (AI2)

Ashish Sabharwal (AI2)

Oren Etzioni
(AI2)

Peter Clark
(AI2)

Shyam Upadhyay
(Uepnn)

Michael Roth
(Saarland Univ)

Thank you!

CogComp-NLP:
`https://github.com/CogComp/cogcomp-nlp`

Questions?