Reasoning-Driven Question Answering for Natural Language Understanding

Daniel Khashabi
Natural Language Understanding
Interpret a given text similar to humans.

Measuring the progress by answering questions.

- A system that is better in understanding language, should have a higher chance of answering these questions.
- This has been used in the field for many years
  [Winograd, 1972; Lehnert, 1977b; others]
  - Question Answering (QA),
  - Reading Comprehension (RC),
  - Textual Entailment (TE), etc.
Interpret a given text similar to humans.

Measuring the progress by answering questions.

- A system that is better in understanding language, should have a higher chance of answering these questions.
- This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
  - Question Answering (QA),
  - Reading Comprehension (RC),
  - Textual Entailment (TE), etc.
Interpret a given text similar to humans.

Measuring the progress by answering questions.
  o A system that is better in understanding language, should have a higher chance of answering these questions.
  o This has been used in the field for many years [Winograd, 1972; Lehnert, 1977b; others]
    • Question Answering (QA),
    • Reading Comprehension (RC),
    • Textual Entailment (TE), etc.
- Interpret a given text similar to humans.

- Measuring the progress by answering questions.
  - A system that is better in understanding language, should have a higher chance of answering these questions.
  - This has been used in the field for many years
    [Winograd, 1972; Lehnert, 1977b; others]
    - Question Answering (QA),
    - Reading Comprehension (RC),
    - Textual Entailment (TE), etc.
Interpret a given text similar to humans.

Measuring the progress by answering questions.
  o A system that is better in understanding language, should have a higher chance of answering these questions.
  o This has been used in the field for many years
    [Winograd, 1972; Lehnert, 1977b; others]
    • Question Answering (QA),
    • Reading Comprehension (RC),
    • Textual Entailment (TE), etc.
NLU Challenges: Ambiguity

- Making sense of strings.
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

A 61-year old furniture salesman
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

A 61-year old furniture salesman
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

A 61-year old furniture salesman

An antique furniture salesman

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

A 61-year old furniture salesman

An antique furniture salesman

The sentence from The New York Times
A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.

- Making sense of strings.
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

- A 61-year old furniture salesman
- An antique furniture salesman
- Shaft of a freight elevator
- Part-whole

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

- A 61-year old furniture salesman
- An antique furniture salesman
- shaft of a freight elevator

The sentence from The New York Times
NLU Challenges: Ambiguity

- Making sense of strings.

"A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday."

- A 61-year old furniture salesman
- An antique furniture salesman
- shaft of a freight elevator
- height of a freight elevator

The sentence from The New York Times
NLU Challenges: Ambiguity

Making sense of strings.

“A 61-year old furniture salesman was pushed down the shaft of a freight elevator yesterday.”

- A 61-year old furniture salesman
- An antique furniture salesman

Oxford English Dictionary lists 10 primary meanings for “of”.

The sentence from The New York Times
NLU Challenges: Variability
“The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter.”
“The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter.”
NLU Challenges: Variability

- A single meaning mentioned in many different ways.

“The buffer springs at the bottom of the shaft prevented the car from crushing the salesman, John J. Hug, after he was pushed from the first floor to the basement. The car stopped about 12 inches above him as he flattened himself at the bottom of the pit. Mr. Hug was pinned in the shaft for about half an hour until his cries attracted the attention of a porter.”

- Even more variability in bigger units (phrases, sentences, paragraphs, etc.)
Lots of understanding is only implied from text.

- the car is significantly heavier than the man;
- he had nowhere to go at the bottom of the pit;
- if he didn’t flatten himself, he would have died;
- ...

The sentence from The New York Times
Lots of understanding is only implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- the car is significantly heavier than the man;
- he had nowhere to go at the bottom of the pit;
- if he didn’t flatten himself, he would have died;
- ...

The sentence from The New York Times
Lots of understanding is only implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- the car is significantly heavier than the man;
- he had nowhere to go at the bottom of the pit;
- if he didn’t flatten himself, he would have died;
- ...

The sentence from The New York Times
Lots of understanding is only implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- The car is significantly heavier than the man;
- He had nowhere to go at the bottom of the pit;
- If he didn’t flatten himself, he would have died;
- ...
Lots of understanding is only implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- the car is significantly heavier than the man;
- he had nowhere to go at the bottom of the pit;
- if he didn’t flatten himself, he would have died;
- ...

The sentence from The New York Times
Lots of understanding is **only** implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- the car is significantly heavier than the man;
- he had nowhere to go at the bottom of the pit;
- if he didn’t flatten himself, he would have died;
- ...

*The sentence from The New York Times*
Lots of understanding is **only** implied from text.

“The car stopped about 12 inches above him as he flattened himself at the bottom of the pit.”

- *the car is significantly heavier than the man;*
- *he had nowhere to go at the bottom of the pit;*
- *if he didn’t flatten himself, he would have died;*
- *...*
NLU Challenges: Small bits, Big conclusions
NLU Challenges: Small bits, Big conclusions

- "Reasoning" as the process of combining facts and beliefs, to make decisions.
  [Johnson–Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of "reasoning":
  - Inductive, deductive, analogy, quantitative, etc.
“Reasoning” as the process of combining facts and beliefs, to make decisions.

[Johnson–Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of “reasoning”:
  - Inductive, deductive, analogy, quantitative, etc.

"... his weeps attracted the attention of a porter"
**NLU Challenges: Small bits, Big conclusions**

- “Reasoning” as the process of combining facts and beliefs, to make decisions.
  
  [Johnson-Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of “reasoning”:
  - Inductive, deductive, analogy, quantitative, etc.

“The story from The New York Times
NLU Challenges: Small bits, Big conclusions

- “Reasoning” as the process of combining facts and beliefs, to make decisions.
  [Johnson–Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of “reasoning”:
  - Inductive, deductive, analogy, quantitative, etc.

“... his weeps attracted the attention of a porter”
NLU Challenges: Small bits, Big conclusions

- “Reasoning” as the process of combining facts and beliefs, to make decisions.
  [Johnson–Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of “reasoning”:
  - Inductive, deductive, analogy, quantitative, etc.

"... his weeps attracted the attention of a porter"

Abductive reasoning

[Peirce, 1883]
NLU Challenges: Small bits, Big conclusions

- “Reasoning” as the process of combining facts and beliefs, to make decisions. [Johnson-Laird, 1980]

- Why was he crying?
  - Maybe he was scared.
  - Maybe he was injured.

- Many other forms of “reasoning”:
  - Inductive, deductive, analogy, quantitative, etc.

“... his weeps attracted the attention of a porter” [Peirce, 1883]

The story from The New York Times
Thesis Structure and Challenges Addressed
Thesis Structure and Challenges Addressed

Approach
Thesis Structure and Challenges Addressed

Approach

1. System design
Thesis Structure and Challenges Addressed

Approach

① System design

② Evaluation
Thesis Structure and Challenges Addressed

Approach

1. System design
2. Evaluation
3. Formalism
Thesis Structure and Challenges Addressed

Challenge
- Variability
- Ambiguity
- Reasoning
- Common-sense understanding

Approach
1. System design
2. Evaluation
3. Formalism
Thesis Structure and Challenges Addressed

Challenge | Variability | Ambiguity | Reasoning | Common-sense understanding
---|---|---|---|---
Approach
1. System design
2. Evaluation
3. Formalism

Abductive reasoning
Thesis Structure and Challenges Addressed

Challenge: Variability, Ambiguity, Reasoning, Common-sense understanding

Approach:
1. System design: Semantic similarity, Semantic abstraction
2. Evaluation: Grounding language
3. Formalism: Abductive reasoning
Thesis Structure and Challenges Addressed

**Challenge**
- Variability
  - Semantic similarity
  - Semantic abstraction
- Ambiguity
  - Grounding language
- Reasoning
  - Abductive reasoning
- Common-sense understanding

**Approach**
1. System design
2. Evaluation
3. Formalism

Reasoning with multiple pieces of text
Thesis Structure and Challenges Addressed

<table>
<thead>
<tr>
<th>Approach</th>
<th>Challenge</th>
<th>Variability</th>
<th>Ambiguity</th>
<th>Reasoning</th>
<th>Common-sense understanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 System design</td>
<td>Semantic similarity</td>
<td>Semantic abstraction</td>
<td>Grounding language</td>
<td>Abductive reasoning</td>
<td>Reasoning with multiple pieces of text</td>
</tr>
<tr>
<td>2 Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reasoning with common-sense understanding</td>
</tr>
<tr>
<td>3 Formalism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thesis Structure and Challenges Addressed

**Approach**

1. **System design**
   - Variability: Semantic similarity, Semantic abstraction
   - Ambiguity: Grounding language
   - Reasoning: Abductive reasoning
   - Common-sense understanding: Reasoning with multiple pieces of text

2. **Evaluation**
   - Variability: Semantic similarity, Semantic abstraction
   - Ambiguity: Grounding language
   - Reasoning: Abductive reasoning
   - Common-sense understanding: Reasoning with common-sense understanding

3. **Formalism**
   - Variability: Semantic similarity, Semantic abstraction
   - Ambiguity: Grounding language
   - Reasoning: Abductive reasoning
   - Common-sense understanding: Reasoning with multiple pieces of text

A framework to study reasoning, in presence of variability and ambiguity.
Thesis Statement

- Progress in **automated** question answering could be facilitated by incorporating the ability to **reason** over natural language **abstractions** and **world knowledge**.

- More **challenging, yet realistic** QA datasets pose problems to current technologies; hence, more **opportunities for improvement**.
Road map

- Introduction and motivation
- Part 1: Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge [IJCAI’16]
  - QA with Semantic Abstractions of Raw Text [AAAI’18]
  - Learning to Pay Attention to Essential Terms in Questions [CoNLL’17]
- Part 2: Moving the Peaks Higher: More Challenging Datasets
  - A QA Benchmark for Reasoning on Multiple Sentences [NAACL’18]
  - A QA Benchmark for Temporal Common-sense [Submitted]
- Part 3: Formal Study of Reasoning in Natural Language
  - Capabilities and Limitations of Reasoning in Natural Language [In submission]
- Conclusion
Road map

- Introduction and motivation

- Part 1: Reasoning-Driven System Design
  - QA as Subgraph Optimization on Tabular Knowledge [IJCAI’16]
  - QA with Semantic Abstractions of Raw Text [AAAI’18]
  - Learning to Pay Attention to Essential Terms in Questions [CoNLL’17]

- Part 2: Moving the Peaks Higher: More Challenging Datasets
  - A QA Benchmark for Reasoning on Multiple Sentences [NAACL’18]
  - A QA Benchmark for Temporal Common-sense [Submitted]

- Part 3: Formal Study of Reasoning in Natural Language
  - Capabilities and Limitations of Reasoning in Natural Language [In submission]

- Conclusion
Part 1: Reasoning-Driven System Design

QA as Subgraph Optimization on Tabular Knowledge \([\text{IJCAI’16}]\)

- Motivation
- Knowledge as Tables
- Reasoning on Knowledge
- Experimental results
QA as Subgraph Optimization on Tabular Internal Knowledge: Overview

[Clark et al, 2015]
Question: In New York State, the longest period of daylight occurs during which month?

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

[Clark et al, 2015]
- Standardized science exams. [Clark et al, 2015]
- Simple language; machines require the ability to use the knowledge and abstract over it.
- The “knowledge” encoded within the solver.

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

**Candidates:** (A) June  (B) March  (C) December  (D) September
Standardized science exams.  [Clark et al, 2015]

Simple language; machines require the ability use the knowledge and abstract over it.

The “knowledge” encoded within the solver.

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

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

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

Premise: a system that “understands” this phenomenon can correctly answer many variations!
In New York State, the longest period of daylight occurs during which month?
(A) June
(B) March
(C) December
(D) September

Premise: a system that “understands” this phenomenon can correctly answer many variations!
In New York State, the longest period of daylight occurs during which month?

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

Premise: a system that “understands” this phenomenon can correctly answer many variations!
In New York State, the longest period of daylight occurs during which month?

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

Premise: a system that “understands” this phenomenon can correctly answer many variations!
Question: In New York State, the longest period of daylight occurs during which month?
(A) June  (B) March  (C) December  (D) September
Semi-Structured Inference

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

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    - regions, hemispheres, solstice, ...
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Semi-Structured Inference

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces:
    - regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Semi-Structured Inference

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

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: *regions, hemispheres, solstice, ....*
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

  ✓ principled approach, explainable answers
  ✓ robust to variations
Structured, Multi-Step Reasoning

- Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
- Reasoning: putting together pieces of knowledge in a principled way.
- Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations

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

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces: *regions, hemispheres, solstice, ...*
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

  ✓ principled approach, explainable answers
  ✓ robust to variations

- New York  
  - Longest Day
    - Northern Hemisphere
      - Summer Solstice
        - month?
          - (A) June
**Semi-Structured Inference**

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

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

- **Diagram:**
  - New Zealand → Longest Day → Northern Hemisphere → Summer Solstice → month?
  - (A) June
Semi-Structured Inference

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

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

- principled approach, explainable answers
- robust to variations

- **New Zealand**
- **Longest Day**
- **Southern Hemisphere**
- **Summer Solstice**
- **New York**
- **Northern Hemisphere**

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

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations

New Zealand
Longest Day
Southern Hemisphere
Summer Solstice
(C) Dec
**Question**: In New York State, the longest period of daylight occurs during which month?
(A) June  (B) March  (C) December  (D) September

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, .....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Semi-Structured Inference

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

- Structured, Multi-Step Reasoning
  - **Science knowledge in small, manageable, swappable pieces:**
    - regions, hemispheres, solstice, ....
  - **Reasoning:** putting together pieces of knowledge in a principled way.
  - **Goal:** overcome brittleness

- ✔ principled approach, explainable answers
- ✔ robust to variations

New Zealand ➔ Southern Hemisphere ➔ Winter Solstice ➔ Shortest Day

(C) Dec
Semi-Structured Inference

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

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces:
    regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Question: In New York State, the longest period of daylight occurs during which month?
(A) June  (B) March  (C) December  (D) September

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces: *regions, hemispheres, solstice, ....*
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Semi-Structured Inference

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

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ...
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
Question: In New York State, the longest period of daylight occurs during which month?
(A) June  (B) March  (C) December  (D) September

- Structured, Multi-Step Reasoning
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness

✓ principled approach, explainable answers
✓ robust to variations
**Semi-Structured Inference**

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

- **Structured, Multi-Step Reasoning**
  - Science knowledge in small, manageable, swappable pieces: regions, hemispheres, solstice, ....
  - Reasoning: putting together pieces of knowledge in a principled way.
  - Goal: overcome brittleness
  
  ✓ principled approach, explainable answers
  ✓ robust to variations

*How can we achieve this?*
Semi-Structured Inference: High-level View

Question Answering as **Global Reasoning** over **Semi-Structured Knowledge**
Question Answering as **Global Reasoning** over **Semi-Structured Knowledge**

- **Knowledge Representation**
  - Structured, yet flexible

**Question Answering**
Semi-Structured Inference: High-level View

Question Answering as **Global Reasoning** over **Semi-Structured Knowledge**
The Necessary Knowledge

The Knowledge Atlas: 12 key sections

Celestial Phenomena
- sun
- moon
- stars
- day/night, rotation
- revolution

The Earth
- air
- water
- land
- weather
- precipitation
- erosion

Matter
- solid/liquid/gas
- properties
- conductivity
- texture
- temperature
- measuring tools

Energy
- forms
- energy transfer
- heat
- electricity
- chemical energy
- energy conversion

Forces
- gravity
- magnetism
- force
- friction
- pull/pushing
- attraction

Living things
- living
- nonliving
- characteristics
- animals
- plants
- fish

Inheritance
- inherited traits
- resemblance
- acquired traits
- learned traits
- body features
- skills

The Environment and Adaptation
- senses
- habitats
- behavior
- camouflage
- survival

Continuity of Life
- life cycle
- life span
- offspring
- reproduction
- coloration
- mating

Life Functions
- breathing
- growing
- eating
- food
- air
- water

Interdependence
- food web
- producers
- consumers
- decomposers
- predators
- prey

Human Impact
- human activities
- environment
- ecosystem
- pollution
- conservation
- deforestation
Knowledge as Frames

Frame Semantics
[Minsky, 1974; Fillmore, 1977]
Knowledge as Frames

Orbital events
Hemisphere: ?
Orbital events: ?
Month: ?

Frame Semantics
[Minsky, 1974; Fillmore, 1977]
Knowledge as Frames

Orbital events
Hemisphere: ?
Orbital events: ?
Month: ?

Frame Semantics
[Minsky, 1974; Fillmore, 1977]
Knowledge as Frames

Frame Semantics

---

Orbital events
- Hemisphere: ?
- Orbital events: ?
- Month: ?

---

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

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

collections of recurring, related, science concepts

---

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

[Frame Semantics: Minsky, 1974; Fillmore, 1977]

[Orbital events: Dalvi et al, 2016]
Knowledge as Frames

Unstructured

- e.g., free form text from books, web
- easy to acquire, difficult to reason with

Structured

- e.g., probabilistic first-order logic rules, ontologies
- "easy" to reason with, difficult to acquire

Frames with free form text

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

Frame Semantics

- Orbital events
  - Hemisphere: ?
  - Orbital events: ?
  - Month: ?

Simple structure, flexible content

- Can acquire knowledge in automated and semi-automated ways

Collections of recurring, related, science concepts

[Frame Semantics: Minsky, 1974; Fillmore, 1977]

[Dalvi et al, 2016]
Road Map

- **Part 1: Reasoning-Driven System Design**
  - QA as Subgraph Optimization on Tabular Knowledge [*IJCAI’16*]
    - Motivation
    - Knowledge as Frames
    - Reasoning on Knowledge
    - Experimental results
TableILP Solver: An Overview

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

TableILP Solver: An Overview

- Question Q with answer options A
- Knowledge Tables T
- Phrasal Entailment
  - Word and short-phrase level entailment / similarity
- Optimization using Integer Linear Program (ILP) formalism

\[
M(T,Q,A) = \max \sum_i c_i x_i \quad \text{subject to} \quad \sum_i a_{1i} x_i \leq b_1, \quad \ldots, \quad \sum_i a_{ki} x_i \leq b_k
\]

\[ x_i \in \mathbb{N} \cup \{0\} \]
A discrete **optimization** approach to QA for multiple-choice questions

**TableILP Solver: An Overview**

- **Sources of knowledge:**
  - Question Q with answer options A
  - Knowledge Tables T

- **ILP model builder**
  - Word and short-phrase level entailment / similarity

- **ILP engine**
  - \( M(T,Q,A) \)
  - \( a_i \in A \) with support graph + score

**Optimization using Integer Linear Program (ILP) formalism**

\[
\begin{align*}
\max & \sum_i c_i x_i \\
\text{s.t.} & \quad \sum_i a_{1i} x_i \leq b_1 \\
& \quad \ldots \\
& \quad \sum_i a_{ki} x_i \leq b_k \\
& \quad x_i \in \mathbb{N} \cup \{0\}
\end{align*}
\]
A discrete optimization approach to QA for multiple-choice questions

TableILP Solver: An Overview

Sources of knowledge:

Question Q with answer options A

Knowledge Tables T

ILP model builder

M(T,Q,A)

ILP engine

a_i ∈ A with support graph + score

Optimization using Integer Linear Program (ILP) formalism

\[
\max \sum_i c_i x_i \\
\forall x_i \in \text{NU}\{0\}
\]

\[
\sum_i a_{1i} x_i \leq b_1 \\
\ldots \\
\sum_i a_{ki} x_i \leq b_k
\]
TableILP Solver: An Overview

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

**Sources of knowledge:**

- Question Q with answer options A
- Knowledge Tables T

**ILP model builder**

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

**ILP engine**

- **a_i ∈ A with support graph + score**

**Optimization using Integer Linear Program (ILP) formalism**

\[
\begin{align*}
\max & \quad \sum_i c_i x_i \\
\text{s.t.} & \quad \sum_i a_{1i} x_i \leq b_1 \\
& \quad \ldots \\
& \quad \sum_i a_{ki} x_i \leq b_k \\
\end{align*}
\]
A discrete optimization approach to QA for multiple-choice questions

TableILP Solver: An Overview

Sources of knowledge:

- Question Q with answer options A
- Knowledge Tables T
- Word and short-phrase level entailment / similarity
- Optimization using Integer Linear Program (ILP) formalism

\[ M(T,Q,A) \]

\[ \max \sum_i c_i x_i \quad \forall x_i \in \text{NU}\{0\} \]

\[ \sum_i a_{1i} x_i \leq b_1 \]

\[ \ldots \]

\[ \sum_i a_{ki} x_i \leq b_k \]
Search for the best **Support Graph** connecting the Question to an Answer through Tables.
Q: In New York State, the longest period of daylight occurs during which month?

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

How is relevant information expressed in my KB?

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

Search for the best **Support Graph** connecting the Question to an Answer through Tables.
Q: In New York State, the longest period of daylight occurs during which month?

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

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

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

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

**Potential Link:**

| Regions and Hemispheres |

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

Search for the best **Support Graph** connecting the Question to an Answer through Tables.
Q: In New York State, the longest period of daylight occurs during which month?

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

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

**TableILP: Main Idea**
**TableILP: Main Idea**

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

<table>
<thead>
<tr>
<th>Subdivision</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>New York State</td>
<td>USA</td>
</tr>
<tr>
<td>California</td>
<td>USA</td>
</tr>
<tr>
<td>Rio de Janeiro</td>
<td>Brazil</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orbital Event</th>
<th>Day Duration</th>
<th>Night Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Summer Solstice</td>
<td>Long</td>
<td>Short</td>
</tr>
<tr>
<td>Winter Solstice</td>
<td>Short</td>
<td>Long</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>North</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td>South</td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

Semi-structured Knowledge

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

Abductive reasoning

[Peirce, 1883]
TableILP: Main Idea

An ideal Support Graph

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

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Hemisphere</th>
<th>Orbital Event</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>Northern</td>
<td>Summer Solstice</td>
<td>June</td>
</tr>
<tr>
<td>Canada</td>
<td>Northern</td>
<td>Winter Solstice</td>
<td>December</td>
</tr>
<tr>
<td>Brazil</td>
<td>Southern</td>
<td>Summer Solstice</td>
<td>December</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Winter Solstice</td>
<td>June</td>
</tr>
</tbody>
</table>

Link this information to identify the best supported answer!

Semi-structured Knowledge

Search for the best Support Graph connecting the Question to an Answer through Tables.
Goal: Design ILP constraints C and objective function F, s.t. maximizing F subject to C yields a “desirable” support graph

- Many possible “proof structures”
  - single/multi-table, single/multi-row, answer in table header, answer spanning multiple cells
- Must balance reward for connections with penalty for spurious links
- Imperfect lexical “similarity” blackbox
- Partial or missing knowledge in tables
- Question logic (negation, conjunction, comparison)
- Scalability of ILP solvers

\[
\begin{align*}
\max & \quad \sum_i c_i x_i \\
\text{s.t.} & \quad \sum_i a_{1i} x_i \leq b_1 \\
& \quad \ldots \\
& \quad \forall x_i \in \mathbb{N}\{0\} \\
& \quad \sum_i a_{ki} x_i \leq b_k
\end{align*}
\]
Operates on lexical units of alignment

- cells + headers of tables T
- question chunks Q
- answer options A

~50 high level constraints + preferences

**Variables** define the space of “support graphs” connecting Q, A, T

- Which nodes + edges between lexical units are active?

**Objective Function:** “better” support graphs = higher objective value

- Reward active units, high lexical match links, column header match, ...
- WH-term boost (“which form of energy...”), science-term boost (“evaporation”)
- Penalize spurious overuse of frequently occurring terms

\[
\max \sum_i c_i x_i \\
\forall x_i \in \mathbb{N}\{0\} \\
\sum_i a_{1i} x_i \leq b_1 \\
\vdots \\
\sum_i a_{ki} x_i \leq b_k
\]
Dual goal: scalability, consider only meaningful support graphs

- **Structural Constraints**
  - Meaningful proof structures
    - connectedness, question coverage, appropriate table use
    - single/multi-table, single/multi-row, etc
  - Simplicity appropriate for 4th / 8th grade

- **Semantic Constraints**
  - Chaining => table joins between semantically similar column pairs
  - Relation matching (ruler measures length, change from water to liquid)
Experimental Results [KKSР'16]
Experimental Results [KKSRS'16]

- Exam Score (%)
  - IR: 58.5
Experimental Results [KKSР’16]

- IR: 58.5%
- PMI: 60.7%
Experimental Results [KKS'16]
Experimental Results [KKSР’16]

Table ILP scores 54% on “IR-hard” questions (random guessing = 25%)
Experimental Results [KKSR’16]

Table ILP scores 54% on “IR-hard” questions (random guessing = 25%)

Ensemble performs 8–10% higher than IR baselines [CEKSTTK’16]
Experimental Results [KSR’16]

<table>
<thead>
<tr>
<th>Exam Score (%)</th>
<th>MLN</th>
<th>IR(tables)</th>
<th>TableILP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>47.5</td>
<td>51.20</td>
<td>61.5</td>
</tr>
</tbody>
</table>
Experimental Results [KKSР’16]

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

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

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

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

(A) *eastern*  (B) June  (C) *history*  (D) *years*
Motivating Example: Circling Back!

- Towards “real understanding” of the phenomenon tested in a question.

| In New York State, the longest period of daylight occurs during which month? | June |
| In New Zealand, the longest period of daylight occurs during which month? | June |
| In New Zealand, the shortest period of daylight occurs during which month? | March |
| In New Zealand, the shortest period of night occurs during which month? | December |
Motivating Example: Circling Back!

- Towards “real understanding” of the phenomenon tested in a question.

<table>
<thead>
<tr>
<th></th>
<th>IR</th>
<th>PMI</th>
<th>TableILP</th>
</tr>
</thead>
<tbody>
<tr>
<td>In New York State, the longest period of daylight occurs during which month?</td>
<td>June</td>
<td>June</td>
<td>June</td>
</tr>
<tr>
<td>In New Zealand, the longest period of daylight occurs during which month?</td>
<td>March</td>
<td>Dec</td>
<td>Dec</td>
</tr>
<tr>
<td>In New Zealand, the shortest period of daylight occurs during which month?</td>
<td>March</td>
<td>Dec</td>
<td>June</td>
</tr>
<tr>
<td>In New Zealand, the shortest period of night occurs during which month?</td>
<td>Dec</td>
<td>Dec</td>
<td>Dec</td>
</tr>
</tbody>
</table>
Summary, So Far

- Elementary school science tests as a challenge for NLU.
- Knowledge as semi-automatically extracted knowledge.
- Abductive reasoning, to provide the best explanations.
- Showed effective and complementary performances.
- Impacts, since publications:
  - Strong performance on new datasets [Clark et al, Arxiv’2018]
  - Inspired other works [Khot et al, ACL’2017]
Part 1: Reasoning-Driven System Design

- QA as Subgraph Optimization on Tabular Knowledge \([IJCAI’16]\)
- QA with Semantic Abstractions of Raw Text \([AAAI’18]\)
- Learning to Pay Attention to Essential Terms in Questions \([CoNLL’17]\)
Reasoning With Semantic Abstractions [KKSR’18]
Reasoning With Semantic Abstractions [KKSR’18]

- Question Q with answer options A
- ILP model builder
- Lexical “similarity”
- Word and short-phrase level entailment / similarity
- Knowledge Tables T
- ILP engine
- $a_i \in A$ with support graph + score
Reasoning With Semantic Abstractions [KKSR‘18]

Knowledge provided as raw text.
Reasoning With Semantic Abstractions [KKSR'18]

Knowledge provided as raw text.

Representing text, as layers of semantic abstractions.
Reasoning With Semantic Abstractions [\textit{KKSR}'18]

- **Verb-Semantic Roles** [Punyakanok et al, 2008]
- **Preposition-Semantic Roles** [Srikumar & Roth, 2013]
- **Comma-semantic Roles** [Arivazhagan et al, 2016]
- **Coreference** [Chang et al, 2012]

Knowledge provided as raw text.

Representing text, as layers of semantic abstractions.

Knowledge Tables $T$

Question $Q$ with answer options $A$

ILP model builder

$M(T, Q, A)$

ILP engine

Word and short-phrase level entailment / similarity

Lexical “similarity”

$a_i \in A$ with support graph + score
Reasoning With Semantic Abstractions [KKSR’18]

Knowledge provided as raw text.

Representing text, as layers of semantic abstractions.

- **Verb–Semantic Roles**  [Punyakanok et al, 2008]
- **Preposition–Semantic Roles**  [Srikumar & Roth, 2013]
- **Comma–semantic Roles**  [Arivazhagan et al, 2016]
- **Coreference**  [Chang et al, 2012]
Reasoning With Semantic Abstractions [KKSR’18]

Knowledge provided as raw text.

Representing text, as layers of semantic abstractions.

- **Verb-Semantic Roles** [Punyakanok et al, 2008]
- **Preposition-Semantic Roles** [Srikumar & Roth, 2013]
- **Comma-semantic Roles** [Arivazhagan et al, 2016]
- **Coreference** [Chang et al, 2012]

Easier domain transfer! Improvements in multiple domains.
Learning Essential Terms in Questions [KKS'R’17]

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

Some *animals* grow *thicker hair* as a season changes. This *adaptation* helps to _______.
(A) find food (B) *keep warmer* (C) grow stronger (D) escape from predators
Learning Essential Terms in Questions [KKSRS’17]

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

Some animals grows thicker hair as a season changes. This adaptation helps to _______.
(A) find food (B) keep warmer (C) grow stronger (D) escape from predators
Learning Essential Terms in Questions [KKS R’17]

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

Some animals grow thicker hair as a season changes. This adaptation helps to _______.
(A) find food (B) keep warmer (C) grow stronger (D) escape from predators

Essentiality in Questions

2K annotated questions
19K annotated terms
Learning Essential Terms in Questions [KKS’17]

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

Some animals grows thicker hair as a season changes. This adaptation helps to _______.
(A) find food (B) keep warmer (C) grow stronger (D) escape from predators

Essentiality in Questions

Important for humans!

2K annotated questions
19K annotated terms
Learning Essential Terms in Questions [KSR’17]

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

Some animals grow thicker hair as a season changes. This adaptation helps to _______.
(A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

Essentiality in Questions

2K annotated questions
19K annotated terms

Important for humans!

State-of-the-art
Essentiality classifier:
F1 = 0.8, MAP = 0.9
Learning Essential Terms in Questions [KKS R’17]

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

Some **animals** grow **thicker hair** as a season changes. This **adaptation** helps to ________. (A) find food (B) **keep warmer** (C) grow stronger (D) escape from predators

---

**Essentiality in Questions**

- 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

**Dataset**
- Baseline
  - Regents: 59.11
  - AI2Public: 57.90
  - RegentsPertd: 61.84
- With ET
  - Regents: 60.85
  - AI2Public: 59.10
  - RegentsPertd: 66.84

**2K annotated questions**

**19K annotated terms**
Road Map

- **Part 2: Moving the Peaks Higher: More Challenging Datasets**
  - A QA Benchmark for Temporal Common Sense [Submitted]
  - A QA Benchmark for Reasoning on Multiple Sentences [NAACL’18]
The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]

- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data
The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar et al, 2016; others]

- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data
The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar at al, 2016; others]

- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data
The Recent State of the Field: Good News

- Many large QA datasets  [Rajpurkar et al., 2016; others]

- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

- Super-human performance in “machine comprehension” (SQuAD)  2018
- Super-human performance in “machine-translation”  2017
- Google deep-net machine translation  2016
- Super-human performance in “object recognition” & “image captioning”  2015
The Recent State of the Field: Good News

- Many large QA datasets [Rajpurkar et al, 2016; others]

- Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attentions)
  - More data

![Timeline of Super-human performance](chart.png)

- Super-human performance in “object recognition” & “image captioning” 2015
- Super-human performance in “machine-comprehension” (SQuAD) 2016
- Google deep-net machine translation 2017
- Super-human performance in “machine-translation” 2018
Many large QA datasets  [Rajpurkar et al., 2016; others]

Successes, with neural nets
  - Faster computers (e.g., GPUs)
  - New computational modules (e.g., Attention)
  - More data
The Recent State of the Field: Bad News
The systems easily break-down. [Kat al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]

- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.

- Discoveries are more about tasks (datasets).

- The urge to scale up datasets has biased in certain angles and limited their diversity.
The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]

- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.

- Discoveries are more about tasks (datasets).

- The urge to scale up datasets has biased in certain angles and limited their diversity.
The Recent State of the Field: Bad News

- The systems easily break-down. [K et al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.
- Discoveries are more about tasks (datasets).
- The urge to scale up datasets has biased in certain angles and limited their diversity.
The Recent State of the Field: Bad News

- The systems easily break-down.  
  [Kat al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]

- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.

- Discoveries are more about tasks (datasets).  
  [Darwiche, 2017]

- The urge to scale up datasets has biased in certain angles and limited their diversity.
The Recent State of the Field: Bad News

- The systems easily break-down. [K at al, 2016; Jia et al, 2017; Belinkov et al, 2018; others]
- Many problems with no significant success: Math word problems; dialogue; many others.
  - Some are not even defined yet.
- Discoveries are more about tasks (datasets). [Darwiche, 2017]
- The urge to scale up datasets has biased in certain angles and limited their diversity.
Moving the Peaks Higher
Goal here:
Define and create challenges not addressed by the community. Challenges that require external knowledge, common-sense, complex reasoning, etc.
Moving the Peaks Higher

Goal here:
Define and create challenges not addressed by the community. Challenges that require external knowledge, common-sense, complex reasoning, etc.

Including our systems
Moving the Peaks Higher

Goal here:
Define and create challenges not addressed by the community. Challenges that require external knowledge, common-sense, complex reasoning, etc.

Including our systems
Road Map

- **Part 2: Moving the Peaks Higher: More Challenging Datasets**
  - A QA Benchmark for Temporal Common Sense *Submitted*
  - A QA Benchmark for Reasoning on Multiple Sentences *NAACL’18*
Did Aristotle have a laptop?

[Valiant, ?]
Did Aristotle have a laptop?

Aristotle's Laptop | Series on Machine Consciousness - World Scientific
https://www.worldscientific.com/worldscibooks/10.1142/8113
This book is about a scientific ingredient that was not available to Aristotle: the science of information. Would the course of the philosophy of the mind have been ...

Aristotle's Laptop: The discovery of out informational mind | Request PDF
https://www.researchgate.net/.../259464884_Aristotle's_Laptop_The_discovery_of_out....
Aristotle's convincing philosophy is likely to have shaped (even indirectly) many ... mind have been different had Aristotle pronounced that the matter of mind was ...

Aristotle's Laptop eBook by Igor Aleksander - 9789814425629 ...
https://www.kobo.com/.../Advanced Computing › Artificial Intelligence
Read "Aristotle's Laptop The Discovery of our Informational Mind" by Igor ... Aristotle's convincing philosophy is likely to have shaped (even indirectly) ... have been different had Aristotle pronounced that the matter of mind was information?
Types of “Knowledge”

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
  - Commonsense knowledge
Types of “Knowledge”

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
    - The capital city of Venezuela is Caracas.
    - Obama was born in Honolulu, HI.
  - Commonsense knowledge
Types of “Knowledge”

- [Loosely-defined] types of knowledge

  - Encyclopedic knowledge
    - The capital city of Venezuela is Caracas.
    - Obama was born in Honolulu, HI.

  - Commonsense knowledge
    - If you lived thousands of years ago, you’re unlikely to be alive now.
    - Laptops didn’t exist, before they were invented.
Types of “Knowledge”

- [Loosely-defined] types of knowledge
  - Encyclopedic knowledge
    - The capital city of Venezuela is Caracas.
    - Obama was born in Honolulu, HI.
  - Commonsense knowledge
    - If you lived thousands of years ago, you’re unlikely to be alive now.
    - Laptops didn’t exist, before they were invented.

Focus of this work
Common Sense: A Short History
Many early works
  - Since the early days of AI
  - Ambitious projects

Recent years:
  - Winograd Schema Challenge
Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects

- Recent years:
  - Winograd Schema Challenge
Common Sense: A Short History

- Many early works
  - Since the early days of AI  [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects  [Lenat, 85; others]

- Recent years:
  - Winograd Schema Challenge
Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]

- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]
Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]

- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]

“Jack pulled up a picture of Aristotle on his laptop”

“Jack pulled up a picture of Aristotle from his biography”
Common Sense: A Short History

- Many early works
  - Since the early days of AI [McCarthy, 68; Charniak, 77; others]
  - Ambitious projects [Lenat, 85; others]

- Recent years:
  - Winograd Schema Challenge [Levesque, 2014; Peng, K, Roth, 2015]

“Jack pulled up a picture of Aristotle on his laptop”

“Jack pulled up a picture of Aristotle from his biography”

Incentivizing commonsense understanding as a high-level and well-defined task.
Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.
Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.

  “Going to barbershop” takes a couple of **hours**
  “Going to college” takes a couple of **years**
Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.

<table>
<thead>
<tr>
<th>Event duration</th>
<th>Event frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Going to barbershop” takes a couple of <strong>hours</strong></td>
<td>“Take a trip to Africa” happens once in a few <strong>years</strong></td>
</tr>
<tr>
<td>“Going to college” takes a couple of <strong>years</strong></td>
<td>“Take a trip to parent’s” happens every few <strong>weeks</strong> or <strong>months</strong></td>
</tr>
</tbody>
</table>
Understanding “time” is a key ability in many NLU tasks.

- “Going to barbershop” takes a couple of **hours**
- “Going to college” takes a couple of **years**
- “Take a trip to Africa” happens once in a few **years**
- “Take a trip to parent’s” happens every few **weeks** or **months**
- “Going to work” usually happens around **morning** time
- “Going to a bar” usually happens around **evening/night** time
Temporal Common Sense

- Understanding “time” is a key ability in many NLU tasks.

<table>
<thead>
<tr>
<th>Event Duration</th>
<th>Event Frequency</th>
<th>Typical Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Going to barbershop” takes a couple of <strong>hours</strong></td>
<td>“Going to college” takes a couple of <strong>years</strong></td>
<td>“Going to work” usually happens around <strong>morning</strong> time</td>
</tr>
<tr>
<td>“Take a trip to Africa” happens once in a few <strong>years</strong></td>
<td>“Take a trip to parent’s” happens every few <strong>weeks</strong> or <strong>months</strong></td>
<td>“Going to a bar” usually happens around <strong>evening/night</strong> time</td>
</tr>
</tbody>
</table>

**Goal of this section:**
QA dataset that requires temporal commonsense.
Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a “temporal” understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: I clapped her shoulder to show I was not laughing at her.

Question: How long did they laugh?

Candidates:
(A) for a few days  (B) 20 minutes  (C) for a few minutes
Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a “temporal” understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: With the amount of money Diggler was making he was able to support both his and Rothchild's addictions.

Question: What time did Diggler go to work?

Candidates:
- (A) at eight in the late night
- (B) he leaves around 3 am
- (C) he leaves around 8 am
Temporal Common Sense as QA

A dataset of natural language questions

- Questions about a “temporal” understanding
  - About 1k questions and 8k candidate answers
  - 5 temporal phenomena:
    - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

**Scenario:** She checked the kitchen, but didn't find anything missing there except for a clock.

**Question:** Why was she checking for missing items?

**Candidates:**
(A) her window had been broken  (B) someone tied her hands
Temporal Common Sense as QA

- A dataset of natural language questions
  - Questions about a “temporal” understanding
    - About 1k questions and 8k candidate answers
    - 5 temporal phenomena:
      - Event duration, event frequency, absolute time point, event ordering, stationary vs transient

Scenario: They then took a boat to Africa and Asia, where they went on a trip through the mountains.

Question: How often do they go on trips?

Candidates:
- (A) every night
- (B) once a year
- (C) twice a year
- (D) once a week
TacoQA: Construction Overview

- Define the task and the desired questions
- Find the qualified annotators
TacoQA: Construction Overview

Define the task and the desired questions

Find the qualified annotators

Generate questions, given scenarios, for each temporal phenomena
TacoQA: Construction Overview

Define the task and the desired questions

Find the qualified annotators

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year
TacoQA: Construction Overview

Define the task and the desired questions

Find the qualified annotators

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?
A: Once a year

Verify the question quality and their temporal category
TacoQA: Construction Overview

1. Define the task and the desired questions
2. Find the qualified annotators
3. Generate questions, given scenarios, for each temporal phenomena
4. Verify the question quality and their temporal category
5. Candidate answer expansion

Q: How often do they go on trips?
A: Once a year

Candidate answers:
- Once a month
- Once a decade
- Twice a year
- Never
- Rarely
TacoQA: Construction Overview

Define the task and the desired questions

Find the qualified annotators

Verify the question quality and their temporal category

Candidate answer expansion

Generate questions, given scenarios, for each temporal phenomena

Q: How often do they go on trips?  
A: Once a year

Answer verification

Candidate answers:
- Once a month
- Once a decade
- Twice a year
- Never
- Rarely
Experimental Results [ZKNR, under review]
Experimental Results [ZKNR, under review]

- Human (F1)
- Human (EM)

- Random
- ESIM
- BERT
Experimental Results [ZKNR, under review]

A system gets credit only if it gets all the candidates right.
Experimental Results [ZKNR, under review]

A system gets credit only if it gets all the candidates right.
A system gets credit only if it gets all the candidates right.

Supervised-learning (LSTM) [Chen et al, 2017]

Experimental Results [ZKKNR, under review]
A system gets credit only if it gets all the candidates right.
Experimental Results [ZKNR, under review]

A system gets credit only if it gets all the candidates right.

Supervised-learning (LSTM)  
[Chen et al, 2017]

Pre-trained contextualized model  
[Devlin et al, 2018]
Experimental Results [ZKNR, under review]

A system gets credit only if it gets all the candidates right.

Supervised-learning (LSTM) [Chen et al, 2017]

Pre-trained contextualized model [Devlin et al, 2018]
Experimental results [ZKNR, under review]

![Graph showing comparison of different models for F1 and EM metrics. The models compared are Random, ESIM, BERT (question+answer), BERT (paragraph+answer), BERT (answer-only), and BERT. The metrics are F1 and EM. The graph shows that BERT generally outperforms the other models.]

[Devlin et al, 2018]
Experimental results [ZKNR, under review]
The State of Current Systems

[Liu & Singh, 2001]

[Devlin et al, 2018; Peters et al, 2018]
The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness

- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues
The State of Current Systems

- **Explicit knowledge bases; e.g. ConceptNet** [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness

- **Language models, soft representations** [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues
The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness

- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

Scenario: He laid down on the chair and pawed at her as she ran in a circle under it.
Question: How long did she run in a circle?
The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness

- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

**Scenario:** He laid down on the chair and pawed at her as she ran in a circle under it.

**Question:** How long did she run in a circle?

---

<table>
<thead>
<tr>
<th>BERT selected as answer</th>
<th>Candidate answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>✘</td>
<td>weeks</td>
</tr>
<tr>
<td>✘</td>
<td>days</td>
</tr>
<tr>
<td>✓</td>
<td>minutes</td>
</tr>
</tbody>
</table>
The State of Current Systems

- Explicit knowledge bases; e.g. ConceptNet [Liu & Singh, 2001]
  - Okay precision but low recall (coverage)
  - Suffer from brittleness

- Language models, soft representations [Devlin et al, 2018; Peters et al, 2018]
  - Can deal with certain associations
  - Suffer from precision issues

**Scenario:** He laid down on the chair and pawed at her as she ran in a circle under it.

**Question:** How long did she run in a circle?

<table>
<thead>
<tr>
<th>BERT selected as answer</th>
<th>Candidate answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>weeks</td>
</tr>
<tr>
<td>✗</td>
<td>days</td>
</tr>
<tr>
<td>✓</td>
<td>minutes</td>
</tr>
<tr>
<td>✓</td>
<td>1 minutes</td>
</tr>
<tr>
<td>✓</td>
<td>28740000 minutes</td>
</tr>
</tbody>
</table>
Summary of This Section

- Understanding time is crucial aspect of NLU.
- A QA dataset of temporal commonsense questions.
- Evaluated systems and showing few angles they are missing.
Road Map

- **Part 2: Moving the Peaks Higher: More Challenging Datasets**
  - A QA Benchmark for Temporal Common-sense [Submitted]
  - A QA Benchmark for Reasoning on Multiple Sentences [NAACL’18]
A Benchmark for Reasoning over Multiple Sentences [KCRUR’18]
“Multi-sentence” hypothesis: Questions that require multiple sentences tend to be “hard”.
The need for creating “reasoning–forcing” challenges

“Multi-sentence” hypothesis: Questions that require multiple sentences tend to be “hard”.

4-step crowdsourcing

From 8 domains (fiction, news, science, etc)

- +10k questions
- 50k candidate answers
- +700 paragraphs

https://cogcomp.org/multirc
A Benchmark for Reasoning over Multiple Sentences [KCRUR’18]

- The need for creating “reasoning-forcing” challenges

  "Multi-sentence" hypothesis:
  Questions that require multiple sentences tend to be “hard”.

- 4-step crowdsourcing

- From 8 domains (fiction, news, science, etc)
  - +10k questions
  - 50k candidate answers
  - +700 paragraphs

https://cogcomp.org/multirc
Part 3: Formal Study of Reasoning in Natural Language

- Capabilities and Limitations of Reasoning in Natural Language [In submission]
A Formal Study of NL Reasoning: Overview
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about “knowledge” and “reasoning”.
  - Information represented as graphs (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like variability, ambiguity, etc.
  - Reasoning: the operation that combines chunks of information to make a conclusion.
- Distinguish successful and failed reasoning.
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about “knowledge” and “reasoning”.
  - Information represented as graphs (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like variability, ambiguity, etc.
  - Reasoning: the operation that combines chunks of information to make a conclusion.
- Distinguish successful and failed reasoning.
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.

- Requires assumptions about “knowledge” and “reasoning”.
  - Information represented as graphs (nodes and semantic relations).
    - Any other structure can be thought of as an explicit or implicit graph.
  - Incorporate properties like variability, ambiguity, etc.
  - Reasoning: the operation that combines chunks of information to make a conclusion.

- Distinguish successful and failed reasoning.
A Formal Study of NL Reasoning: Overview

- We provide a formalized study of reasoning.
- Requires assumptions about “knowledge” and “reasoning”.
  - Information represented as graphs (nodes and semantic relations).
    - Any other structure can be thought of an explicit or implicit graph.
  - Incorporate properties like variability, ambiguity, etc.
  - Reasoning: the operation that combines chunks of information to make a conclusion.
- Distinguish successful and failed reasoning.
We provide a formalized study of reasoning.

Requires assumptions about “knowledge” and “reasoning”.

- Information represented as graphs (nodes and semantic relations).
  - Any other structure can be thought of an explicit or implicit graph.
- Incorporate properties like variability, ambiguity, etc.
- Reasoning: the operation that combines chunks of information to make a conclusion.

Distinguish successful and failed reasoning.
What We Do Not Say
What We Do Not Say

- **Not** making claims about:
  - How “reasoning” should be defined.
  - How “knowledge” should be represented.
  - How systems should be designed.

- **Theoretical results based on assumptions** (“no free lunch”).
  - Which may or may not stand the test of time.
What We Do Not Say

- **Not** making claims about:
  - How “reasoning” should be defined.
  - How “knowledge” should be represented.
  - How systems should be designed.

- Theoretical results based on assumptions ("no free lunch").
  - Which may or may not stand the test of time.
Tale of Two Spaces

Meaning Graph

- Conceptualization
- No ambiguity
- No variability
- No missing relations
Tale of Two Spaces

Meaning Graph
- Conceptualization
- No ambiguity
- No variability
- No missing relations
Tale of Two Spaces

Symbol Graph
- Physical world
- Ambiguity
- Variability
- Missing relations

Utterance

Meaning Graph
- Conceptualization
- No ambiguity
- No variability
- No missing relations
Tale of Two Spaces

Symbol Graph
- Physical world
- Ambiguity
- Variability
- Missing relations

Utterance

Meaning Graph
- Conceptualization
- No ambiguity
- No variability
- No missing relations
Formalizing Symbol and Meaning Space

Symbol Graph

Meaning Graph

CPU

Chips
Formalizing Symbol and Meaning Space
Formalizing Symbol and Meaning Space
Formalizing Symbol and Meaning Space

Symbol Graph
- IC
- Chip
- CPU
- Central Processing Unit
- Potato Chips

Meaning Graph
- Utterance
- CPU
- Chips

222
Formalizing Symbol and Meaning Space

**Symbol Graph**
- IC
- Chip
- CPU
- Central Processing Unit
- Potato Chips

**Meaning Graph**
- Utterance
- CPU
- Chips

Variability
Formalizing Symbol and Meaning Space

Symbol Graph
- IC
- Chip
- Chips
- CPU
- Central Processing Unit
- Potato Chips

Variability

Meaning Graph
- Utterance
- CPU
- Chips

Variability
Formalizing Symbol and Meaning Space

Symbol Graph
- IC
- Chip
- Chips
- CPU
- Central Processing Unit
- Potato Chips

Meaning Graph
- Utterance
- CPU
- Chips

Variability
Ambiguity
Formalizing Symbol and Meaning Space
Formalizing Symbol and Meaning Space

Symbol Graph

- Missing relations
- Variability
- Central Processing Unit
- Chips
- Potato Chips
- IC
- Chip

Meaning Graph

- Utterance
- Ambiguity

IC Chip Chips
Formalizing Symbol and Meaning Space

Symbol Graph

Missing relations

IC

CPU

Central Processing Unit

Chip

Chips

Potato Chips

Spurious relations

Variability

Ambiguity

Meaning Graph

Utterance

Missing relations

Variability

Spurious relations
Formalizing Symbol and Meaning Space

**Symbol Graph**

- **Missing relations**
  - $\rho_+$: Probability of retaining relations

- **Variability**
  - $\lambda$: variability factor

- **Spurious relations**
  - $p_-$: Probability of adding spurious relations

**Meaning Graph**

- **Utterance**
  - $\varepsilon_+$: Probability of inferring true equivalence

- **Ambiguity**
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")

**Symbol Graph**
- "present day spoons"
- "metal spoon"
- "stainless-metals"
- "metals"

**Meaning Graph**
- "high thermal conductivity"
- "good conductors of heat"
- "present day spoons are made from metal such as steel"
- "Metals in general have high electrical conductivity, and high thermal conductivity."
- "dense materials such as metals and stone are good conductors of heat"

**Utterance**
- Q: has-property(metal-spoon, thermal-conductor)
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")

Q: "is a metal spoon a good conductor of heat?"

Q: has-property(metal-spoon, thermal-conductor)
Reasoning by Combining Local Information

- Reasoning itself is hard to define.
- Class of reasoning which functions by combining local information ("multi-hop")

Q: "is a metal spoon a good conductor of heat?"

Q: has-property(metal-spoon, thermal-conductor)
The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)
The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)

Symbol graph
The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)
The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)

![Symbol graph](image1)

\[ \text{Utterance} \]

![Meaning graphs](image2)

- one of the hypotheses require d-step connectivity
The Inference Problem

- “Inferring” connectivity in the (hidden) meaning graph
  - Given observations (a symbol graph)

**Goal:** Infer the connectivity of two given nodes (in the unseen meaning graph), given observations in the symbol graph.
Results: Big picture [KS KSR, in submission]
Results: Big picture [KSRSK, in submission]
Results: Big picture \([K\text{SKSR, in submission}]\)
Results: Big picture [KSKSR, in submission]
Results: Big picture [SKSR, in submission]

\[ p \]  

Noise parameter

\[ n \]  

Number of steps required

\[ d \]
Theorem A (informal)

For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$. 

Number of steps required

Noise parameter
Theorem A (informal)
For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$. 
Theorem A (informal)
For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$.

Theorem B (informal)
If $d \in \Omega(\log n)$ and “sufficient” noise, no algorithm can confidently distinguish $H_1$ and $H_2$. 
Results: Big picture [KSKR, in submission]

Theorem A (informal)
For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$.

Theorem B (informal)
If $d \in \Omega(\log n)$ and “sufficient” noise, no algorithm can confidently distinguish $H_1$ and $H_2$. 
**Theorem A**

For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$.

**Theorem B**

If $d \in \Omega(\log n)$ and “sufficient” noise, no algorithm can confidently distinguish $H_1$ and $H_2$. 

**Results: Big picture** [KSKSR, in submission]

![Diagram](diagram.png)
Theorem A (informal)
For any $d$ there is a choice of parameters such that an algorithm can confidently distinguish $H_1$ and $H_2$.

Theorem B (informal)
If $d \in \Omega(\log n)$ and “sufficient” noise, no algorithm can confidently distinguish $H_1$ and $H_2$. 
Pursuing “very long” multi-hop reasoning is unlikely to result in general results.

Corollary: one has to focus on richer representations (i.e., dealing with **ambiguity** and **variability**) such that it leads to few number of hops needed.
Summary of this section

▪ A framework for studying “reasoning”, in the context of language problems.

▪ Multi-hop reasoning:
  ▪ + There are non-trivial problems where successful reasoning is reliable.
  ▪ − Reasoning with “large”-many hops likely to fail, even with small amount of noise.

▪ Implications for practice
  o Hypothesis: invest in representations that lead to few hops reasonings.
Summary of the talk
Summary of the talk

- NLU; potentials for significant impacts in the coming years.

- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

- Approaches:
  - **System design**: systems that abstracting over text and reasoning with it.
  - **Evaluation**: effective benchmarks to measure and incentivize the community.
  - **Formalism**: to study a class of reasoning algorithms in the context of language.
Summary of the talk

- NLU; potentials for significant impacts in the coming years.

- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
Summary of the talk

- NLU; potentials for significant impacts in the coming years.

- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
Summary of the talk

- NLU; potentials for significant impacts in the coming years.

- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
Summary of the talk

- NLU; potentials for significant impacts in the coming years.
- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.
- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
NLU; potentials for significant impacts in the coming years.

Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
Summary of the talk

- NLU; potentials for significant impacts in the coming years.

- Answering questions: a natural evaluation protocol.
  - Many challenges along the way to this goal: ambiguity, variability, etc.

- Approaches:
  - System design: systems that abstracting over text and reasoning with it.
  - Evaluation: effective benchmarks to measure and incentivize the community.
  - Formalism: to study a class of reasoning algorithms in the context of language.
Thesis Publication

- KSKR. On the Capabilities and Limitations of Reasoning for Natural Language Understanding, in submission.

- ZKNR. A Question Answering Benchmark for Temporal Common-sense, under review.

- KCRUR. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences, NAACL, 2018.


Other Publications

- **NLP:**
  - CWKR. *Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims*, under review.
  - ZKCR. *Zero-Shot Open Entity Typing as Type-Compatible Grounding*, EMNLP, 2018.

- **NLP software/tools:**

- **ML/Optimization/etc:**
  - KSKCSSR. *Relational Learning and Feature Extraction by Querying over Heterogeneous Information Networks*, StartAI, 2018.
  - QK. *Online Learning with Adversarial Delays*, NourIPS, 2015.
Dan Roth (UPenn)
Tushar Khot (AI2)
Ashish Sabharwal (AI2)
Snigdha Chaturvedi (UCSC)
Michael Roth (Saarland U)
Shyam Upadhyay (Upenn → Google)
Erfan Sadeqi Azer (Indiana U)
Ben Zhou (UIUC → UPenn)
Oren Etzioni (AI2)
Peter Clark (AI2)
That’s it folks