Multi-hop Inference for Sentence-Level TextGraphs: How Challenging is Meaningfully Combining Information for Science QA?

Peter Jansen, School of Information, University of Arizona

cognitiveai.org
My long term interest is in building inference algorithms capable of answering questions and producing human-readable explanations by aggregating information from multiple sources and knowledge bases.
My long term interest is in building inference algorithms capable of answering questions and producing human-readable explanations by aggregating information from multiple sources and knowledge bases.

Three observations in the context of science exams and multi-hop inference:

- Standardized elementary science exam questions require an average of 4 to 6 facts (range 1-16) to answer and explain their reasoning.  
  *(Jansen et al., COLING 2016; Jansen et al., LREC 2018 Explanation Corpus)*

- Assembling long chains of facts to answer questions is challenging. Inference algorithms typically struggle to aggregate more than 2 pieces of information together due to “semantic drift”.  
  *(Fried et al., TACL 2015; Khashabi et al., IJCAI 2016; Jansen et al., CL 2017)*

- Fundamental questions about multi-hop inference, like empirically characterizing the difficulty of building graphs of n-facts, are critical to understanding performance, and building better models.  
  *(This work)*
Why Generate Explanations?

**Applications:** Explanations are required for usability in certain high-impact domains where the cost of making errors is high. *(Medicine, scientific inference, intelligent tutoring)*

**Generalization:** Recent work in the cognition and pedagogy literature suggests that explanations (when being tutored) and self-explanations (when engaged in self-directed learning) are an important aspect of the learning process, and help us generalize the knowledge we have learned. *(Roscoe and Chi, 2007; Legare, 2014; Rittle-Johnson and Loehr, 2016)*

**Human-like thought:** Generating compelling human-readable explanations adds constraints to the kinds of inference that can be used, arguably meaning we must develop methods that are more human cognition-like than we might otherwise.
Science Exams Require Complex Inference

Standardized science exams require many challenging forms of complex reasoning to solve, including causal, process, and model-based reasoning.

(Clark et. al, AKBC 2013, Jansen et al., COLING 2016)

Complex Inference (77%)

- Domain-specific models (39%)
- Definitions (12%)
- Properties of Objects (9%)
- Examples of Situations (18%)
- Causality (11%)
- Processes (9%)

Retrieval Methods (23%)

- Information Retrieval Search
- Word Matching
- Information Aggregation
- World Knowledge
- Causal Inference
- Cognition?
Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone  
(B) A girl eating an apple  
(C) An insect crawling on a leaf  
(D) A boy planting tomatoes
Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone  
(B) A girl eating an apple  
(C) An insect crawling on a leaf  
(D) A boy planting tomatoes

Rarely will we be able to retrieve a single passage in a corpus that directly answers a given question:

“"A girl eating an apple is an example of an organism taking in nutrients..."
Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes

- **Girl**: "a girl means a human girl"; "humans are living organisms" (Simple Wiktionary)
- **Eating**: "eating is when an organism takes in nutrients in the form of food" (4th Grade Study Guide)
- **Apple**: "an apple is a kind of fruit"; "fruits are foods" (Simple Wiktionary)
Q: Which of the following is an example of an organism taking in nutrients?

(A) A dog burying a bone
(B) A girl eating an apple
(C) An insect crawling on a leaf
(D) A boy planting tomatoes
Q: Which of the following organisms is a producer?

(A) Frog  (C) Grass
(B) Mushroom  (D) Lizard

Model Answer: (C) Grass

Explanation: A producer is an organism that produces its own food, and is food for other organisms: usually a green plant. Grass is a green, leafy plant that covers the ground.
# Analysis: Explanation Quality

<table>
<thead>
<tr>
<th>Explanation Rating</th>
<th>Information Retrieval Baseline</th>
<th>TAG Information Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Explanation</td>
<td>45%</td>
<td>57%</td>
</tr>
<tr>
<td>Partial Explanation</td>
<td>35%</td>
<td>21%</td>
</tr>
<tr>
<td>Topical Explanation</td>
<td>12%</td>
<td>14%</td>
</tr>
<tr>
<td>Offtopic Explanation</td>
<td>8%</td>
<td>8%</td>
</tr>
</tbody>
</table>

Analysis includes all correctly answered questions by each model across a corpus of 1,000 elementary science questions.
Results: Accuracy by Sentences in TAG

- 1S+2S+3S+IR Ensemble
- 1S+2S Ensemble

Information Aggregation Gain (+5%)

Precision at 1

- 35%
- 40%
- 45%

1 Sentence Graphs
2 Sentence Graphs
3 Sentence Graphs

Semantic Drift (-1.5%)
Nodes represent sentences from a corpus
Edges represent two sentences having lexical overlap (sharing one or more of the same words)
Nodes represent sentences from a corpus
Edges represent two sentences having lexical overlap (sharing one or more of the same words)
Nodes represent sentences from a corpus
Edges represent two sentences having lexical overlap (sharing one or more of the same words)
For a given node in the graph... what proportion of traversals would be meaningful? (i.e. help make a compelling explanation for a given science exam question)

Nodes represent sentences from a corpus
Edges represent two sentences having lexical overlap (sharing one or more of the same words)
Methods
Methods

Data
Drew 14 representative science questions from AI2 question repository. (changes of state, planetary motion, adaptations, life cycle, magnetism, etc.)

Sentence Collection
Collected all sentences from a corpus that had lexical overlap with the question or answer terms.

Manual Ratings
Through 9,784 manual judgements, rated the perceived utility of each sentence towards explaining why a given answer is correct on a 4-point scale.

Analysis
Analyzed these ratings in different scenarios to determine chance performance at assembling 1, 2, or 3 sentence graphs for each corpus.
Corpora

Explanations Corpus (Artificial In-domain)

- 1,365 sentences constructed specifically to build detailed explanations for 400 science exam questions.

Study Guides (In-domain)

- 2,503 sentences from two study guides for the New York and Virginia science exams.

Simple Wikipedia (Near-domain)

- 848,920 sentences from Simple Wikipedia included in the Aristo Mini corpus, flagged as loosely being relevant to the science domain.
Q: What is the main purpose of the flowers on a peach tree?
A: to attract bees for pollination

Highly Likely
The flower helps the plant reproduce because it contains the pollen and eggs.

Possible
Seeds grow in the center of a flower and continue to develop there after the petals fall off the plant.

Topical/Unlikely
There are four major parts of a plant: roots, stem, leaves, and flower.

Offtopic
The average life span of a worker bee is 1 year.
Q: Which of the following organisms is a producer?  
A: Grass

What is chance performance at being able to build 1, 2, or 3 sentence paths that contain sentences that are highly-relevant to building an explanation for the question?

S1: An example of an organism that is also a producer is grass.

S1: A producer is an organism that makes its own food, like green plants.  
S2: Grass is a green leafy plant that covers the ground.

S1: A producer is an organism that makes its own food.  
S2: Photosynthesis allows green leafy plants to make their own food.  
S3: Grass is a green leafy plant that covers the ground.
Q: Which of the following *organisms* is a *producer*?

A: *Grass*

What is *chance performance* at being able to build 1, 2, or 3 sentence paths that contain sentences that are *highly-relevant* to building an explanation for the question?

1 sentence case: *No multi-hop inference*

*S1*: A *producer* is an *organism* that makes its own food, like *green plants*.

*S2*: *Grass* is a *green* leafy *plant* that covers the ground.

3 sentence case: *Sparse*, even in million-sentence corpora

*S2*: Photosynthesis allows *green leafy plants* to make their own food.

*S3*: *Grass* is a *green leafy plant* that covers the ground.
Analysis
## Analysis 1: Likelihood of sentences being relevant

What proportion of sentences with lexical overlap to the question or answer contain highly relevant information?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>1,364 sents</th>
<th>2,503 sents</th>
<th>848,920 sents</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanation Corpus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highly Likely</td>
<td>5.5%</td>
<td>18.4%</td>
<td></td>
</tr>
<tr>
<td>Possible</td>
<td>4.8%</td>
<td>8.5%</td>
<td></td>
</tr>
<tr>
<td><strong>Study Guide Corpus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highly Likely</td>
<td>1.7%</td>
<td>5.2%</td>
<td></td>
</tr>
<tr>
<td>Possible</td>
<td>2.1%</td>
<td>6.3%</td>
<td></td>
</tr>
<tr>
<td><strong>Simple Wikipedia Corpus</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Highly Likely</td>
<td>0.1%</td>
<td>0.5%</td>
<td></td>
</tr>
<tr>
<td>Possible</td>
<td>0.2%</td>
<td>1.1%</td>
<td></td>
</tr>
</tbody>
</table>
### Analysis 1: Likelihood of sentences being relevant

What proportion of sentences with lexical overlap to the question or answer contain highly relevant information?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sent(Q)</th>
<th>Sent(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation Corpus (1,364 sents)</td>
<td>5.5%</td>
<td>18.4%</td>
</tr>
<tr>
<td></td>
<td>4.8%</td>
<td>8.5%</td>
</tr>
<tr>
<td>Study Guide Corpus (2,503 sents)</td>
<td>1.7%</td>
<td>5.2%</td>
</tr>
<tr>
<td></td>
<td>2.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Simple Wikipedia Corpus (848,920 sents)</td>
<td>0.1%</td>
<td>0.5%</td>
</tr>
<tr>
<td></td>
<td>0.2%</td>
<td>1.1%</td>
</tr>
</tbody>
</table>
Analysis 2: Likelihood of 2 sentence aggregations

Probability of selecting 2 sentences that are highly relevant to the question: 1 sent w/Q terms, 1 sent w/A terms, and both sents share other words.

S1: A producer is an organism that makes its own food, like green plants.
S2: Grass is a green leafy plant that covers the ground.

Explanation Corpus (1,364 sents)
- Highly Likely 13.4%
- Highly Likely + Possible 21.0%

Study Guide Corpus (2,503 sents)
- Highly Likely 3.0%
- Highly Likely + Possible 6.6%

Simple Wikipedia Corpus (848,920 sents)
- Highly Likely 0.04%
- Highly Likely + Possible 0.3%
Analysis 2: Likelihood of 2 sentence aggregations

Probability of selecting 2 sentences that are highly relevant to the question: 1 sent w/Q terms, 1 sent w/A terms, and both sents share other words.

**S1:** A producer is an organism that makes its own food, like green plants.

**S2:** Grass is a green leafy plant that covers the ground.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Likely</th>
<th>Likely + Possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation Corpus (1,364 sents)</td>
<td>13.4%</td>
<td>21.0%</td>
</tr>
<tr>
<td>Study Guide Corpus (2,503 sents)</td>
<td>3.0%</td>
<td>Remarkably High</td>
</tr>
<tr>
<td>Simple Wikipedia Corpus (848,920 sents)</td>
<td>0.04%</td>
<td>Extremely Low</td>
</tr>
</tbody>
</table>
Analysis 3: Performance by # of Overlapping Words

The chance likelihood of meaningfully combining 2 sentences from a corpus increases **dramatically** as the number of “other” (not in the Q or A) words shared by those sentences increases.

- Exceeds the best reported multi-hop improvements in the literature (~ +4%)
- Comes at the expense of sparsity.
- Not possible to have high overlap for all Qs
Information aggregation – combining sentences to make inferences and explanations – is **very hard**.

This paper provides the **first empirical measure of chance performance on the multi-hop inference task** using sentences from 3 corpora.

Standardized elementary science exam questions require an average of **4 to 6 facts (range 1-16)** to answer and explain their reasoning.  
*(Jansen et al., COLING 2016; Jansen et a., LREC 2018 Explanation Corpus)*

Chance performance at meaningfully combining 2 sentences ranges from **3-6% (Study Guides)** to **0.04 - 0.3% (Simple Wikipedia)**  
*(This work)*

But combining facts is exponentially difficult. Chance performance in the general case at generating **real 4 fact explanations** is approximately **1 in 200,000 (Study Guides)** to **1 in 17 million (Simple Wikipedia)**.
BUT... chance performance at combining 2 sentences is remarkably high, and can be extremely high in certain scenarios (e.g. many overlapping words).

This is troubling, because generally we’ve only demonstrated successfully combining 2 sentences for QA/explanation generation, and chance performance of the knowledge resources may be very near reported performance of complex systems.

Methods Recommendations:

Evaluate Random Performance of knowledge resources under conditions/features being tested.

Hard: Requires large manual evaluation, unclear that this can be done through crowdsourcing.

Evaluate and Report Model Performance as # of hops increases.

Easy: Trivial to evaluate, but currently treated as a hyperparameter. Transparently shows performance on inference chains of different size, and allows comparing vs chance performance.
Some types of trees are able to survive the heat of a forest fire.

Which of the following characteristics would best help a tree survive a fire?

*C* thick bark

*BG* protecting something means preventing harm to that something

*LG* damage means harm

*CE* as the thickness (of something; of an object) increases, the resistance to damage (of that object; of the)

*CE* fire causes harm (to trees; to forests; to living things)

*CE* bark is a protective covering around the (trunk of; branches of) a tree

*GR* protecting a living thing has a positive impact on that living thing's (survival; health)

*CE* protection means resistance to damage increases

*CE* thickness is a measure of how thick an object is

*GR* a tree is a kind of living thing

*CE* thickness is a property of an object and includes ordered values of (thin; thick)

*CE* bark is a part of a tree

*LG* a part of something means a characteristic of something

*LG* helping something has a positive impact on that something
Nodes represent questions
Edges represent 2+ shared sentences/facts in explanations
Thank You!

Download expanded paper, data, and explanations at: cognitiveai.org/explanationbank

I’m looking for a Postdoctoral Fellow to join the lab (visit cognitiveai.org for contact information)