

# What's Missing: A Knowledge Gap Guided Approach for Multi-hop Question Answering

Tushar Khot and Ashish Sabharwal and Peter Clark  
EMNLP 2019

---

Krunal Shah (shahkr@seas.upenn.edu)  
20<sup>th</sup> April 2020

# Example (from Open Book Question Answering)

---

**Question:** Which of these would let the most **heat travel through**?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

**Knowledge Gap** (similar gaps for other choices):

steel spoon in a cafeteria \_\_\_\_\_ **metal**.

# Motivation

---

- This work concentrates on multi-hop QA in partial knowledge setting
- Most existing QA datasets assume availability of all required knowledge
- However, partial knowledge setting is quite intuitive, natural and challenging

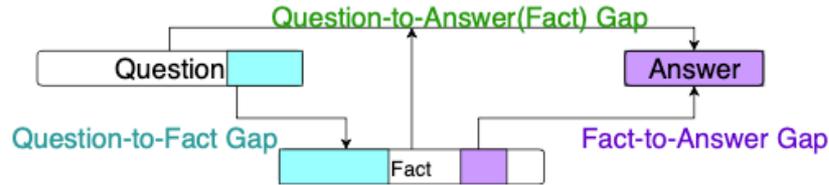
## Example:

Which of these would let the most heat travel through?

a) Osmium

Hints?

# Knowledge Gaps



Broadly (inexhaustively) classified into three types:

1. Question-to-fact Gap:
2. Fact-to-Answer Gap (this work)
3. Question-to-Answer (Fact) Gap

Question: What can cause Kool-aid to disappear ?  
Fact: If a liquid disappears then that liquid probably evaporated  
Answer: Evaporation  
Gap: Kool-aid is a liquid

Question: What can cause liquid to disappear ?  
Fact: If a liquid disappears then that liquid probably evaporated  
Answer: Heat  
Gap: Heat causes evaporation

Question: What is the satellite of the blue planet ?  
Fact: The blue planet refers to planet Earth  
Answer: Moon  
Gap: Moon is the satellite of Earth

# Brief Approach

**Question:** Which of these would let the most **heat travel through**?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

I. Identify “key span” in core fact

# Brief Approach

**Question:** Which of these would let the most **heat travel through**?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

2. Retrieve some knowledge to help fill knowledge gap

# Brief Approach

**Question:** Which of these would let the most **heat travel through**?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

3. Identify relations between key span  
and answer choices  
(using retrieved knowledge)

(steel spoon, **isA**, metal)

# Open Book QA-Short

Narrow down questions as:

1. Fact supports correct answer
  - answer-fact gap
2. Questions with small answer choices
  - Long answer choices lead to noisy gaps

Question: What can cause liquid to disappear ?

Fact: If a liquid disappears then that liquid probably evaporated

Answer: Heat

Gap: Heat causes evaporation

# Knowledge Gap Dataset (KGD)

Starting with questions included in Open Book QA-Short:

## Annotate:

1. Key span in core fact that **could answer the question**  
- because answer-fact gap
2. One or more relations that satisfy knowledge gap

Data point: {question, fact, spans, relations}

Only include questions with agreement greater than 2/3.

Question	Fact	Span	Relation	Gap
<b>Q:</b> A light bulb turns on when it receives energy from <b>A:</b> gasoline	a light bulb converts <i>electrical energy</i> into light energy when it is turned on	electrical energy	provides <sup>-1</sup> , enables <sup>-1</sup>	(gasoline, provides, electrical energy)
<b>Q:</b> What makes the best wiring? <b>A:</b> Tungsten	wiring requires an <i>electrical conductor</i>	electrical conductor	isa <sup>-1</sup> , madeof	(Tungsten, is an, electrical conductor)

# Proposed Model (GapQA)

---

**Question:** Which of these would let the most heat travel through?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

**Core Fact:**

Metal lets heat travel through.

# Proposed Model (GapQA)

**Question:** Which of these would let the most **heat travel through**?

- A) a new pair of jeans.
- B) a steel spoon in a cafeteria.
- C) a cotton candy at a store.
- D) a calvin klein cotton hat.

Repeat subsequent steps for all options

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

# Proposed Model (GapQA)

**Question:** Which of these would let the most **heat travel through**?

A) a new pair of jeans.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

I. Run RC Model with (Question, Core Fact as context) to predict “key span”

# Proposed Model (GapQA)

**Question:** Which of these would let the most **heat travel through**?

A) a new pair of jeans.

**Core Fact:** (annotated as part of dataset)

**Metal** lets **heat travel through**.

## 2. Knowledge Retrieval

Tuple Search:

Subject matching  $s$

Object matching  $c$

Text Search:

Elastic Search using “ $s + c$ ”

Convert everything to text.

# Proposed Model (GapQA)

**Question:** Which of these would let the most **heat travel through**?

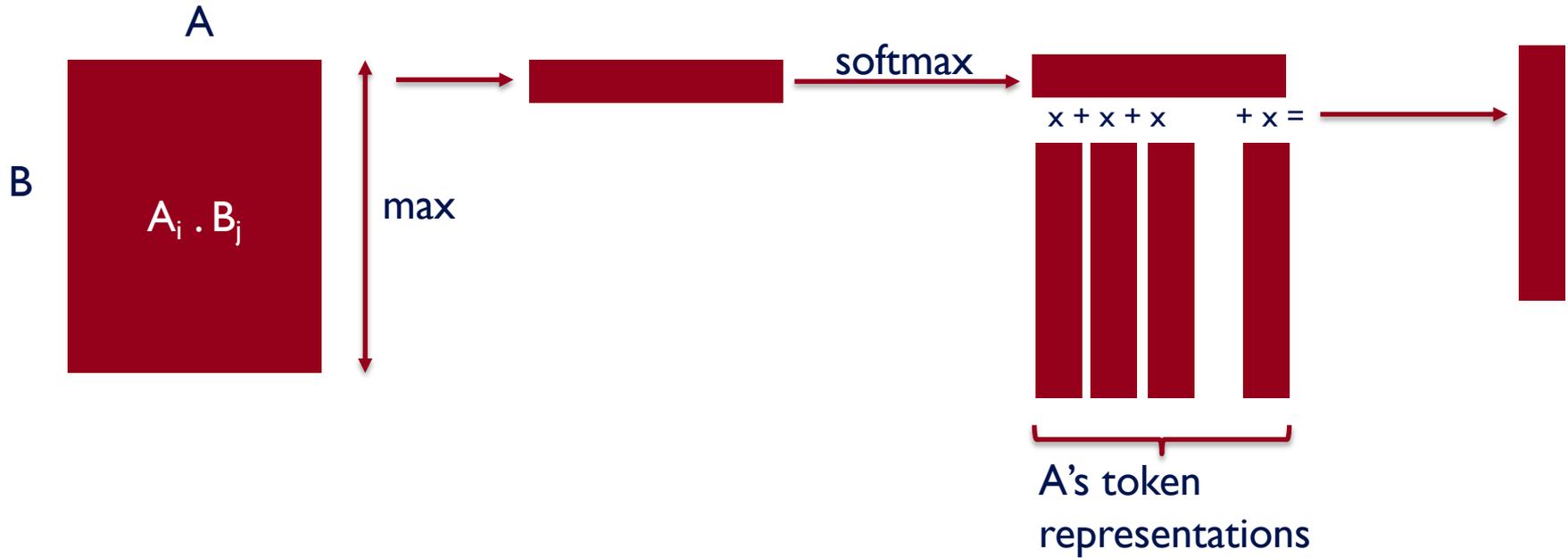
A) a new pair of jeans.

3. Scores the option choice
  - Fact relevance score:
  - Relation prediction score

**Core Fact:** (annotated as part of dataset)

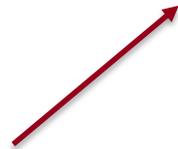
**Metal** lets **heat travel through**.

# B weighted representation of A ( $\text{att}_B(A)$ )



# Compare rep A and rep B (comp(A, B))

representation A



representation B



# Fact Relevance

Question and choice attended fact representation:

$$S_{\text{question, choice}}(\text{fact}) = \text{avg}(\text{att}_{\text{question}}(\text{fact}), \text{att}_{\text{choice}}(\text{fact}))$$

Aggregate fact representation:

$$R_{\text{fact}} = \text{avg}(\text{fact representations})$$

$$\text{score}_{\text{fact}}(\text{choice}) = \text{FF}(\text{comp}(R_{\text{fact}}, S_{\text{question, choice}}(\text{fact})) )$$

# Relation Prediction

Span attended knowledge sentence representation:

$$S_{\text{span}}(\text{knowledge sent}) = \text{att}_{\text{span}}(\text{knowledge sent})$$

Choice attended knowledge sentence representation:

$$S_{\text{choice}}(\text{knowledge sent}) = \text{att}_{\text{choice}}(\text{knowledge sent})$$

Partial relation representation

$$= \text{FF}(\text{comp}(S_{\text{span}}(\text{knowledge sent}), S_{\text{choice}}(\text{knowledge sent})) )$$

# Relation Prediction Score

## Relation Representation

$$R_{\text{relation}} = \text{avg}_{\text{knowledge sentences}}(\text{Partial Relation representation})$$

## Question Fact Composed Representation

$$R_{\text{fact,question}} = \text{comp}(\text{max pooled question rep.}, \text{max pooled fact rep.})$$

## Relation Prediction Score

$$= \text{FF}([R_{\text{fact,question}}; R_{\text{relation}}])$$

# Training methodology

---

1. Use BiDAF trained on SQuAD + finetuned on Knowledge Gap Dataset to predict spans
2. Relation loss:
  - Project relation representation to multilabel relation classification
  - Binary cross-entropy loss
3. Train model on
  - Knowledge Gap Dataset
  - Open Book QA-Short using predicted spans and ignoring relation loss

# Results and analysis

Impressive results:

~6.5% improvement in partial knowledge setting

~3% improvement on complete knowledge setting

Model	OBQA-Short	OBQA-Full
Q2Choice	47.10 ± 1.5	49.64 ± 1.3
ESIM + ELMo	45.93 ± 2.6	49.96 ± 2.5
KER (only f)	57.93 ± 1.4	55.80 ± 1.8
KER (f + WordNet)	54.83 ± 2.5	55.84 ± 1.7
KER (f + OMCS)	49.65 ± 2.0	52.50 ± 0.8
GapQA (f + KB) [Ours]	<b>64.41 ± 1.8*</b>	<b>59.40 ± 1.3*</b>

\* denotes results are statistically significant 21

# Results and analysis

## Ablation study results

Model	OBQA-Short	$\Delta$
GapQA	<b><math>64.41 \pm 1.8</math></b>	—
No Annotations	$58.90 \pm 1.9$	5.51
Heuristic Span Anns.	$61.38 \pm 1.5$	3.03
No Relation Score	$60.48 \pm 1.1$	3.93
No Spans (Model)	$62.14 \pm 2.1$	2.27
No Spans (IR)	$61.79 \pm 1.0$	2.62

# Personal thoughts

---

## About the paper:

- + Propose an excellent (realistic) task
- + Thorough analysis of results
- Only handle two hop questions
  - Jansen et al. show that even elementary science questions require 4 to 6 sentences to answer and explain on an average <sup>[3]</sup>.
- Need a more general framework for dealing with gaps
  - Handle more kinds of knowledge gap
- Results and limitations of the work suggest we are still quite far away from truly solving two hop questions

# Personal thoughts

- The ability to answer the following is **integral** to **reasoning**  
“What more do I need to know (to achieve something)?”
- An observation that can possibly help towards a more general framework

Question: What can cause Kool-aid to disappear ?  
Fact: If a liquid disappears then that liquid probably evaporated  
Answer: Evaporation  
Gap: Kool-aid is a liquid

Question: What can cause liquid to disappear ?  
Fact: If a liquid disappears then that liquid probably evaporated  
Answer: Heat  
Gap: Heat causes evaporation

Question: What is the satellite of the blue planet ?  
Fact: The blue planet refers to planet Earth  
Answer: Moon  
Gap: Moon is the satellite of Earth

- Previously explored idea of multi-hop QA as path finding! [4][5]
- Recognize where a node is missing

# References

---

1. What's Missing: A Knowledge Gap Guided Approach for Multi-hop Question Answering. Tushar Khot, Ashish Sabharwal, Peter Clark. EMNLP 2019
2. QASC: A Dataset for Question Answering via Sentence Composition. Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, Ashish Sabharwal. AACL 2020
3. What's in an Explanation? Characterizing Knowledge and Inference Requirements for Elementary Science Exams. Jansen, Peter and Balasubramanian, Niranjan and Surdeanu, Mihai and Clark, Peter. COLING 2016.
4. Question Answering via Integer Programming over Semi-Structured Knowledge. Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, Dan Roth. IJCAI 2016.
5. Question Answering as Global Reasoning over Semantic Abstractions. Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Dan Roth. AACL 2018.

# Results and analysis

Analyzing results using different knowledge sources

Knowledge Source	Model	OBQA-Short
f + WordNet	KER	54.83 ± 2.5
	GapQA	<b>60.69 ± 1.1*</b>
f + OMCS	KER	49.65 ± 2.0
	GapQA	<b>60.90 ± 2.4*</b>
f + CN + ARC	GapQA	<b>64.41 ± 1.8</b>

\* denotes results are statistically significant 26

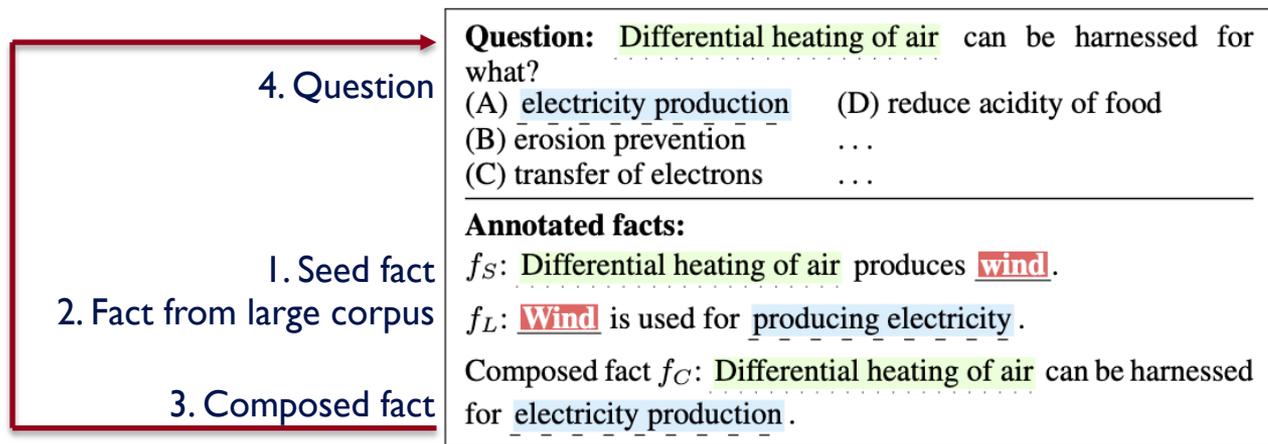
# Contributions

---

- Annotate a subset of Open Book QA for partial knowledge setting
- Propose a novel approach of identify missing knowledge and filling it for multi hop QA
- Propose a model that learns to fill missing knowledge from external knowledge and compose it with existing knowledge
- State of the art results on QA with partial knowledge

# Another related dataset

## QASC: Question Answering via Sentence Composition [2]



# Outline:

---

- Problem and Motivation
- Brief overview and dataset
- Proposed model (GapQA)
- Experiments
- Personal thoughts