A Annotating Gaps: Data Collection

We first identify relevant facts for questions and then collect annotations for fact-answer gap, given the relevant fact. However, straightforward approaches to annotate all questions led to noisy labels. To improve annotation quality, we identified question subsets most suitable for this task and split the fact-answer gap annotation into two steps.

Fact Relevance. The OpenBookQA dataset provides the core science fact used to create the question. However, in 20% of the cases, while the core science fact inspired the question, it is not needed to answer the question (Mihaylov et al., 2018). We also noticed that often multiple facts from the open book can be relevant for a question. So we first create an annotation task to identify the relevant facts from a set of retrieved facts. Also to ensure that there is a gap between the fact and the correct answer, we select facts that have no word overlap with the correct choice or have overlap with multiple answer choices. This ensures that the fact alone can not be trivially used to answer the question.

We ask Turkers to annotate these retrieved facts as (1) are they *relevant* to the question and (2) if relevant, do they point to a *unique* answer. We introduced the second category after noticing that some generic facts can be relevant but not point to a specific answer making identifying the knowledge gap impossible. E.g. The fact: "evaporation is a stage in the water cycle process" only eliminates one answer option from "The only stage of the water cycle process that is nonexistent is (A) evaporation (B) evaluation (C) precipitation (D) condensation". For each question, we selected facts that were marked as relevant and unique by at least two out of three turkers.

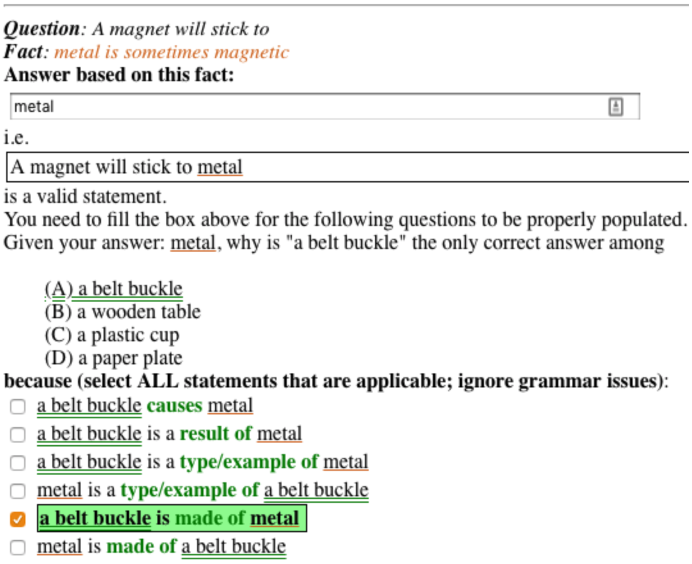
Knowledge Gap. In the second round of data collection, we asked Turkers to write the facts connecting the relevant fact to the correct answer choice. We restricted this task to Masters level Turkers with 95% approval rating and 5000 approved hits. However, we noticed that crowdsource workers would often re-state part of the knowledge mentioned in the original fact or directly connect the question to the answer. This issue was also mentioned by the authors of OpenBookQA who also noticed that the additional facts were "*noisy (incomplete, over-complete, or only distantly related)*" (Mihaylov et al., 2018). E.g. for the question: "In the desert, a hawk may enjoy an occasional (A) coyote (B) reptile (C) bat (D) scorpion" and core fact: "hawks eat lizards", one of the turk-authored additional fact: "Hawks hunt reptiles which live in the desert" is sufficient to answer the question on its own.

We also noticed that questions with long answer choices often have multiple fact-answer gaps leading to complex annotations, e.g. "tracking time" *helps with* "measuring how many marshmallows I can eat in 10 minutes". Collecting knowledge gaps for such questions and common-sense knowledge to capture these gaps are interesting directions of future research. We instead focus on questions where the answer choices have at most two non-stopword tokens. We refer to this subset of questions in OpenBookQA as *OBQA-Short*, which still forms more than 50% of the OpenBookQA set. This subset also forms the target question set of our approach.

Further to simplify this task, we broke the task of identifying the required knowledge into two steps (shown in Figure 7 in Appendix): (1) identify key terms in the core fact that could answer the question, and (2) identify the relationship between these terms and the correct answer choice. For key terms, we asked the Turkers to select spans from the core fact itself, to the extent possible. For the relation identification, we provided a list of relations and asked them to select all the relations that hold between the key term and the correct choice but do not hold for the incorrect answer choices. Based on our analysis, we picked nine most common relations: {causes, definedAs, enables, isa, located in, made of, part of, provides, synonym of} and their inverses (except synonymy).[16] If none of these relations were valid, they were allowed to enter the relation in a text box.

We note that the goal of this effort was to collect supervision for a subset of questions to guide the model and show the value of minimal annotation on this task. We believe our approach can be useful to collect annotations on other question sets as well, or can be used to create a challenge dataset for this sub-task. Moreover, the process of collecting this data revealed potential issues with collecting annotations for knowledge gaps and also inspired the design of our two-step QA model.

[16] These relations were also found to be important by prior approaches (Clark et al., 2014; Khashabi et al., 2016; Jansen et al., 2016, 2018) in the science domain.



Question: *A magnet will stick to*
Fact: *metal is sometimes magnetic*
Answer based on this fact:

metal

i.e.

A magnet will stick to metal

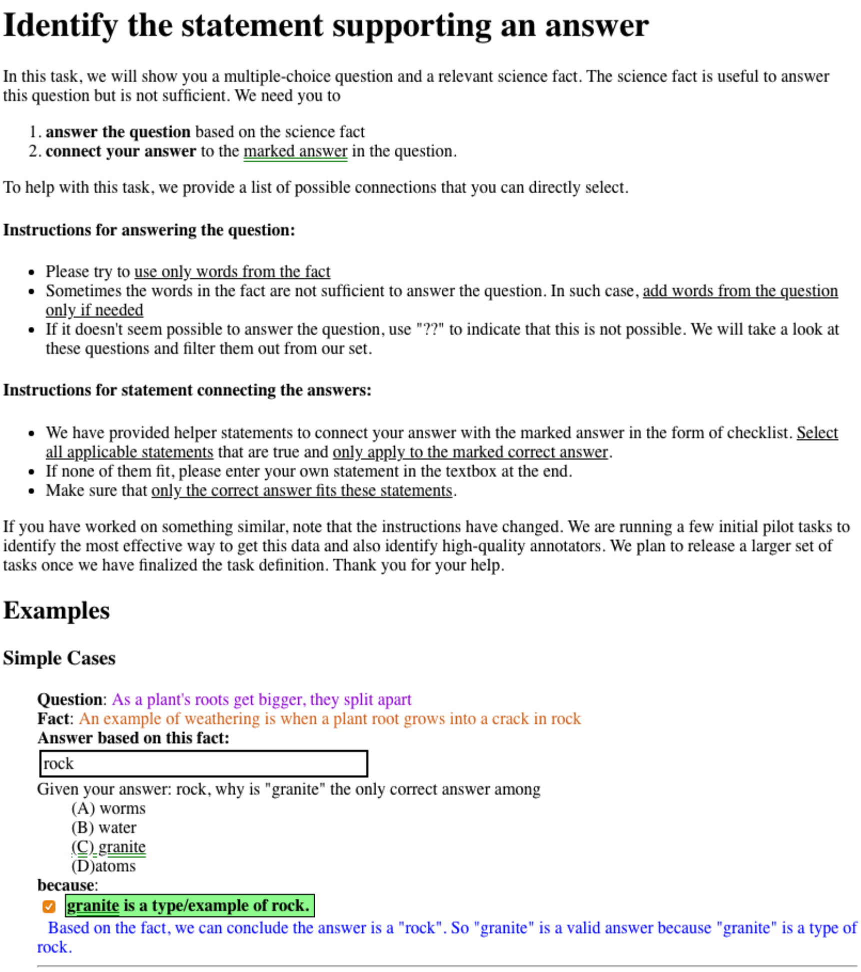
is a valid statement.
You need to fill the box above for the following questions to be properly populated.
Given your answer: metal, why is "a belt buckle" the only correct answer among

(A) a belt buckle
(B) a wooden table
(C) a plastic cup
(D) a paper plate

because (select ALL statements that are applicable; ignore grammar issues):

- □ a belt buckle **causes** metal
- □ a belt buckle is a **result of** metal
- □ a belt buckle is a **type/example of** metal
- □ metal is a **type/example of** a belt buckle
- ☑ **a belt buckle is made of metal**
- □ metal is **made of** a belt buckle

Figure 7: Interface provided to Turkers to annotate the missing fact. Entering the answer span from the fact, metal, in this example, automatically populates the interface with appropriate statements. The valid statements are selected by Turkers and capture the knowledge gap.



Identify the statement supporting an answer

In this task, we will show you a multiple-choice question and a relevant science fact. The science fact is useful to answer this question but is not sufficient. We need you to

1. **answer the question** based on the science fact
2. **connect your answer** to the marked answer in the question.

To help with this task, we provide a list of possible connections that you can directly select.

Instructions for answering the question:

- Please try to use only words from the fact
- Sometimes the words in the fact are not sufficient to answer the question. In such case, add words from the question only if needed
- If it doesn't seem possible to answer the question, use "??" to indicate that this is not possible. We will take a look at these questions and filter them out from our set.

Instructions for statement connecting the answers:

- We have provided helper statements to connect your answer with the marked answer in the form of checklist. Select all applicable statements that are true and only apply to the marked correct answer.
- If none of them fit, please enter your own statement in the textbox at the end.
- Make sure that only the correct answer fits these statements.

If you have worked on something similar, note that the instructions have changed. We are running a few initial pilot tasks to identify the most effective way to get this data and also identify high-quality annotators. We plan to release a larger set of tasks once we have finalized the task definition. Thank you for your help.

Examples

Simple Cases

Question: As a plant's roots get bigger, they split apart
Fact: An example of weathering is when a plant root grows into a crack in rock
Answer based on this fact:

rock

Given your answer: rock, why is "granite" the only correct answer among

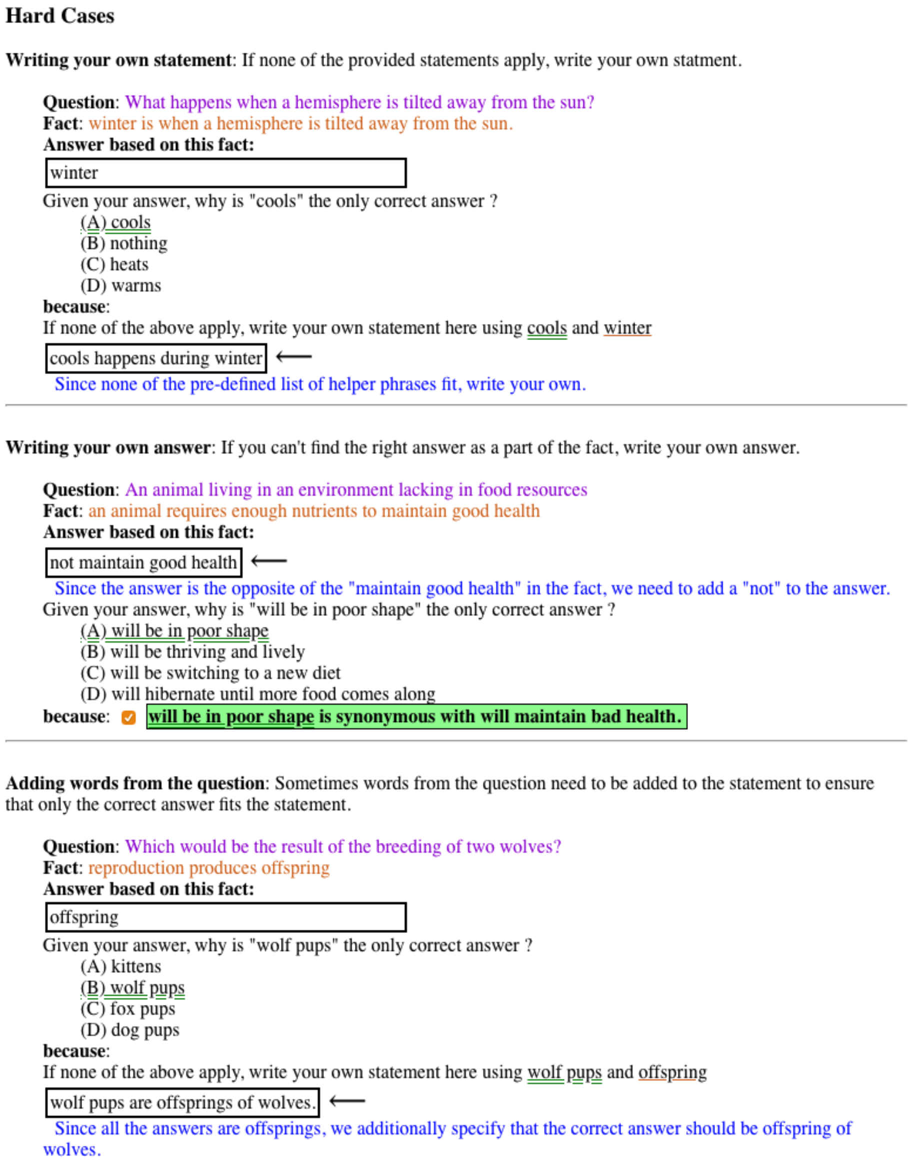
(A) worms
(B) water
(C) granite
(D)atoms

because:

☑ **granite is a type/example of rock.**

Based on the fact, we can conclude the answer is a "rock". So "granite" is a valid answer because "granite" is a type of rock.

Figure 8: Basic Instructions for the task



Hard Cases

Writing your own statement: If none of the provided statements apply, write your own statment.

Question: What happens when a hemisphere is tilted away from the sun?
Fact: winter is when a hemisphere is tilted away from the sun.
Answer based on this fact:

winter

Given your answer, why is "cools" the only correct answer ?

(A) cools
(B) nothing
(C) heats
(D) warms

because:
If none of the above apply, write your own statement here using cools and winter

cools happens during winter ⟵

Since none of the pre-defined list of helper phrases fit, write your own.

Writing your own answer: If you can't find the right answer as a part of the fact, write your own answer.

Question: An animal living in an environment lacking in food resources
Fact: an animal requires enough nutrients to maintain good health
Answer based on this fact:

not maintain good health ⟵

Since the answer is the opposite of the "maintain good health" in the fact, we need to add a "not" to the answer.

Given your answer, why is "will be in poor shape" the only correct answer ?

(A) will be in poor shape
(B) will be thriving and lively
(C) will be switching to a new diet
(D) will hibernate until more food comes along

because: ☑ **will be in poor shape is synonymous with will maintain bad health.**

Adding words from the question: Sometimes words from the question need to be added to the statement to ensure that only the correct answer fits the statement.

Question: Which would be the result of the breeding of two wolves?
Fact: reproduction produces offspring
Answer based on this fact:

offspring

Given your answer, why is "wolf pups" the only correct answer ?

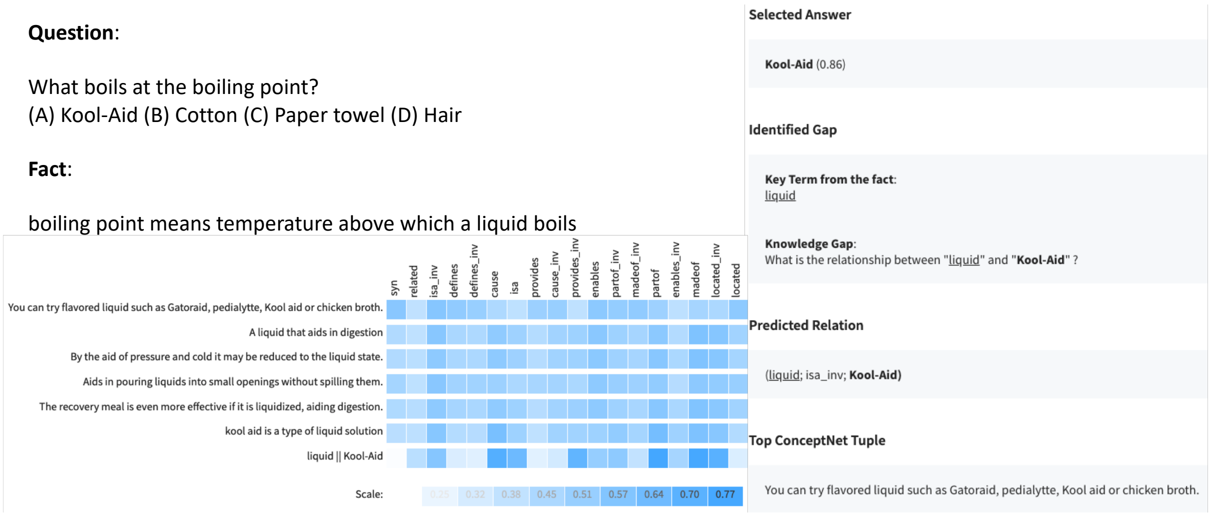
(A) kittens
(B) wolf pups
(C) fox pups
(D) dog pups

because:
If none of the above apply, write your own statement here using wolf pups and offspring

wolf pups are offsprings of wolves. ⟵

Since all the answers are offsprings, we additionally specify that the correct answer should be offspring of wolves.

Figure 9: Instructions for complex examples



Question:

What boils at the boiling point?
(A) Kool-Aid (B) Cotton (C) Paper towel (D) Hair

Fact:

boiling point means temperature above which a liquid boils

syn
related
isa_inv
defines
defines_inv
cause
isa
provides
cause_inv
provides_inv
enables
partof_inv
madeof_inv
partof
enables_inv
madeof
located_inv
located
You can try flavored liquid such as Gatoraid, pedialytte, Kool aid or chicken broth.
A liquid that aids in digestion
By the aid of pressure and cold it may be reduced to the liquid state.
Aids in pouring liquids into small openings without spilling them.
The recovery meal is even more effective if it is liquidized, aiding digestion.
kool aid is a type of liquid solution
liquid || Kool-Aid
Scale:
0.25
0.32
0.38
0.45
0.51
0.57
0.64
0.70
0.77

Selected Answer

Kool-Aid (0.86)

Identified Gap

Key Term from the fact:
liquid

Knowledge Gap:
What is the relationship between "liquid" and "**Kool-Aid**" ?

Predicted Relation

(liquid; isa_inv; **Kool-Aid**)

Top ConceptNet Tuple

You can try flavored liquid such as Gatoraid, pedialytte, Kool aid or chicken broth.

Figure 10: Visualization of the models behavior with the predicted span, top predicted relation, and the top fact used by model. The heat map shows the confidence of the model for all the relations for each input sentence (first five) and ConceptNet sentencized tuple (last but one) and the back-off tuple (last one) to capture the knowledge in the embeddings.

B Implementation Details

We implement all our models in Pytorch (Paszke et al., 2017) using the AllenNLP (Gardner et al., 2017) toolkit. We also used the AllenNLP implementation of the BiDAF model for span prediction. We use 300D 840B Glove (Pennington et al., 2014) embeddings and use 200 dimensional hidden representations for the BiLSTM shared between all inputs (each direction uses 100 dimensional hidden vectors). We use 100 dimensional representations for the relation prediction, $\mathcal{R}_j$. Each feedforward network, FF is a 2-layer network with relu activation, 0.5 dropout (Srivastava et al., 2014), 200 hidden dimensions on the first layer and no dropout on the output layer with linear activation. We use a variational dropout (Gal and Ghahramani, 2016) of 0.2 in all the BiLSTMs. The relation prediction loss is scaled by $\lambda = 1$. We used the Adam (Kingma and Ba, 2015) optimization with initial $lr = 0.001$ and a learning rate scheduler that halves the learning rate after 5 epochs of no change in QA accuracy. We tuned the hyper-parameters and performed early stopping based on question answering accuracy on the validation set. Specifically, we considered {50, 100, 200} dimensional representations, $\lambda \in \{0.1, 1, 10\}$, retrieving {10, 20} knowledge tuples and {[x - y; x*y], [x, y]} combination functions for $\bigotimes$ during the development of the model. The baseline models were developed for this dataset using hyper-parameter tuning; we do not perform any additional tuning. Our model code and pre-trained models are available at https://github.com/allenai/missing-fact.

C ConceptNet sentences

Given a tuple $t = (s, v, o)$, the sentence form is generated as "s is $\text{split}(v)$ o" where $\text{split}(v)$ splits the ConceptNet relation v into a phrase based on its camel-case notation. For example, (belt buckle, /r/MadeOf, metal) would be converted into "belt buckle is made of metal".

D Text retrieval

For each span $\hat{s}$ and answer choice c_i, we query an ElasticSearch [17] index on the input text corpus with the "$\hat{s}$ + c_i" as the query. We also require the matched sentence must contain both the span and the answer choice. We filter long sentences (>300 characters), sentences with negation and noisy sentences[18] from the retrieved sentences.

[17] https://www.elastic.co/

[18] Sentences are considered clean if they contain alphanumeric characters with standard punctuation, start with an alphabet or a number, are single sentence and only uses hyphens in hyphenated word pairs