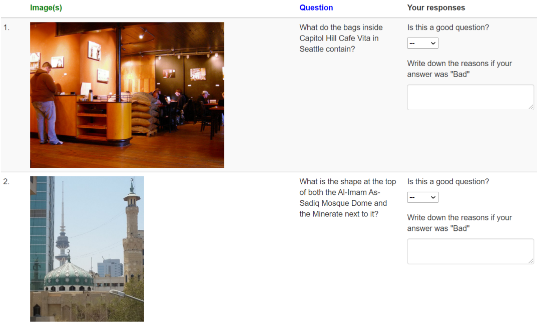
WebQA: Multihop and Multimodal QA Supplementary Material

Yingshan Chang[1] **Mridu Narang**[2] **Hisami Suzuki**[2]
Guihong Cao[2] **Jianfeng Gao**[3] **Yonatan Bisk**[1,3]
[1]Carnegie Mellon University [2]Microsoft, Bing Search [3]Microsoft Research

A. Data Annotation Details

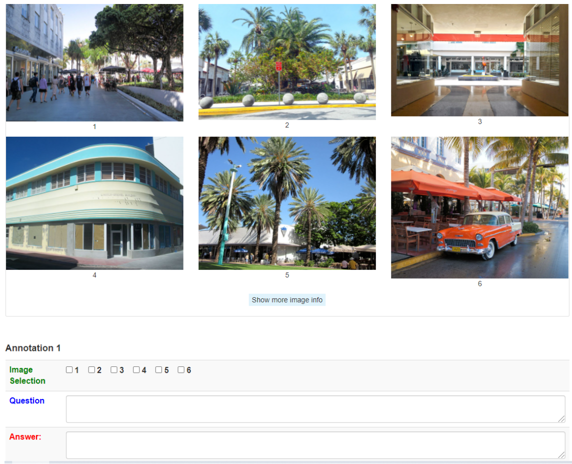
Qualification HIT. For quality control, we included a qualification task with 15 hard coded QA pair annotations, some of which obviously violate the annotation guidelines. Annotators had to point out the problematic pairs and explain in what ways they did not follow the instructions. We restricted to crowdworkers located in the US or Canada, with a general requirement of over 1,000 previously approved HITs with at least 95% approval rate. Additionally, one has to score 80% or higher on our qualification task before getting access to our main task. We gave workers who achieved 60% - 80% at their first attempt a second chance because we believe that workers who had the patience to complete their first attempt were more coachable than others.



	Image(s)	Question	Your responses
1.		What do the bags inside Capitol Hill Cafe Vita in Seattle contain?	Is this a good question? -- Write down the reasons if your answer was "Bad"
2.		What is the shape at the top of both the Al-Imam As-Sadiq Mosque Dome and the Minerate next to it?	Is this a good question? -- Write down the reasons if your answer was "Bad"

Image Filter HIT. We designed a Filter HIT as a pre-step to obtain groups of related images as prompts for the QA-pair creation task. We present 10 images at a time, which are returned by an Image Search API call using the same search term. Annotators were told to a) select 3 out of the 10 that are distinct but related in some ways, and b) give a label that best summarizes the commonality. After having all these image triples, we paired up triples to form groups 6 according to the cosine similarity between their topic labels. We tuned similarity thresholds to make sure that within each group all images fall under the same topic but still have enough dissimilarity to facilitate both connection-based and comparison-based QA-pair construction.

QA Pair Creation HIT. The main annotation task (QA-pair creation task) was released batchwise. We spot checked data quality after every batch and sent targeted feedback when we noticed any deviation from our expectations. Workers who constantly failed to follow the guidelines were de-qualified. Crowdsourcing data is challenging in that crowdworders are usually income-driven and will stick to a fixed answer generation pattern once they find it lucrative. To better align the crowdworkers' incentives with our goal, we gave generous bonuses to the annotations that demonstrate out-of-the-box thinking.

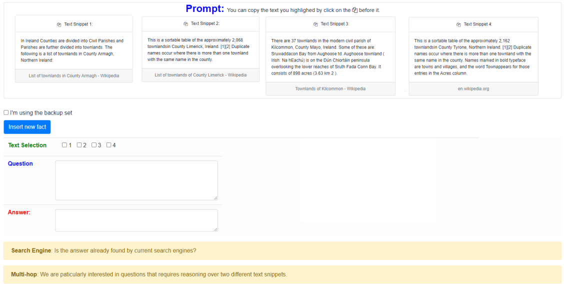


Annotation 1

Image Selection ☐1 ☐2 ☐3 ☐4 ☐5 ☐6

Question

Answer:



Prompt: You can copy the text you highlighed by click on the ⧉ before it.

⧉ Text Snippet 1:	⧉ Text Snippet 2:	⧉ Text Snippet 3:	⧉ Text Snippet 4:
In Ireland Counties are divided into Civil Parishes and Parishes are further divided into townlands. The following is a list of townlands in County Armagh, Northern Ireland:	This is a sortable table of the approximately 2,068 townlandsin County Limerick, Ireland. [1][2] Duplicate names occur where there is more than one townland with the same name in the county.	There are 37 townlands in the modern civil parish of Kilcommon, County Mayo, Ireland. Some of these are: Sruwaddacon Bay from Aughoose td. Aughoose townland (Irish: Na hEachú) is on the Dún Chiortáin peninsula overlooking the lower reaches of Sruth Fada Conn Bay. It consists of 898 acres (3.63 km 2).	This is a sortable table of the approximately 2,162 townlandsin County Tyrone, Northern Ireland. [1][2] Duplicate names occur where there is more than one townland with the same name in the county. Names marked in bold typeface are towns and villages, and the word Townappears for those entries in the Acres column.
List of townlands in County Armagh - Wikipedia	List of townlands of County Limerick - Wikipedia	Townlands of Kilcommon - Wikipedia	en.wikipedia.org

☐ I'm using the backup set

Text Selection ☐1 ☐2 ☐3 ☐4

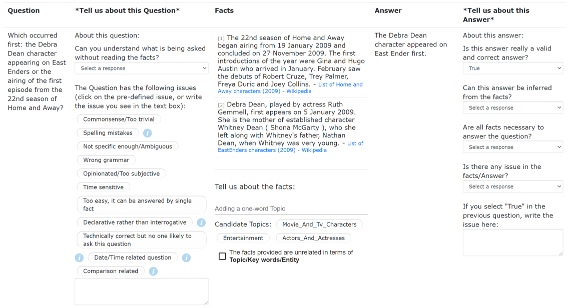
Question

Answer:

Search Engine: Is the answer already found by current search engines?

Multi-hop: We are paticularly interested in questions that requires reasoning over two different text snippets.

QA Pair Validation HIT.



Question

Which occurred first: the Debra Dean character appearing on East Enders or the airing of the first episode from the 22nd season of Home and Away?

Tell us about this Question

About this question:

Can you understand what is being asked without reading the facts?

Select a response

The Question has the following issues (click on the pre-defined issue, or write the issue you see in the text box):

Commonsense/Too trivial

Spelling mistakes

Not specific enough/Ambiguous

Wrong grammar

Opinionated/Too subjective

Time sensitive

Too easy, it can be answered by single fact

Declarative rather than interrogative

Technically correct but no one likely to ask this question

Date/Time related question

Comparison related

Facts

[1] The 22nd season of Home and Away began airing from 19 January 2009 and concluded on 27 November 2009. The first introductions of the year were Gina and Hugo Austin who arrived in January. February saw the debuts of Robert Cruze, Trey Palmer, Freya Duric and Joey Collins. - List of Home and Away characters (2009) - Wikipedia

[2] Debra Dean, played by actress Ruth Gemmell, first appears on 5 January 2009. She is the mother of established character Whitney Dean (Shona McGarty), who she left along with Whitney's father, Nathan Dean, when Whitney was very young. - List of EastEnders characters (2009) - Wikipedia

Tell us about the facts:

Adding a one-word Topic

Candidate Topics: Movie_And_Tv_Characters Entertainment Actors_And_Actresses

☐ The facts provided are unrelated in terms of **Topic/Key words/Entity**

Answer

The Debra Dean character appeared on East Ender first.

Tell us about this Answer

About this answer:

Is this answer really a valid and correct answer?

True

Can this answer be inferred from the facts?

Select a response

Are all facts necessary to answer the question?

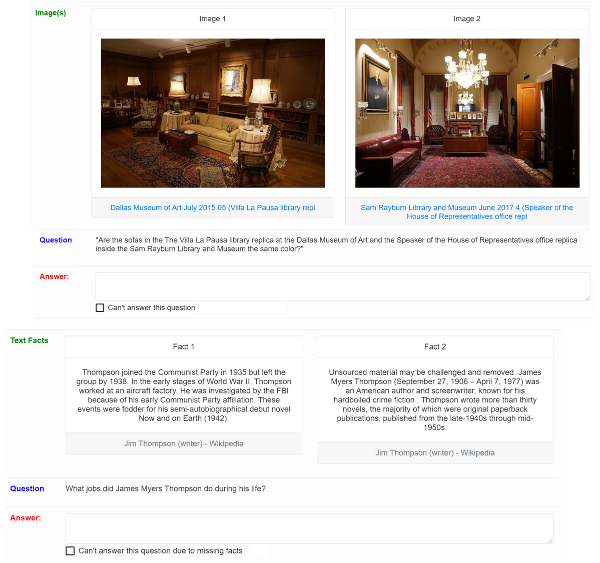
Select a response

Is there any issue in the facts/Answer?

Select a response

If you select "True" in the previous question, write the issue here:

Multiple Human References Generation HIT.



Image(s)

Image 1

Dallas Museum of Art July 2015 05 (Villa La Pausa library repl

Image 2

Sam Rayburn Library and Museum June 2017 4 (Speaker of the House of Representatives office repl

Question "Are the sofas in the The Villa La Pausa library replica at the Dallas Museum of Art and the Speaker of the House of Representatives office replica inside the Sam Rayburn Library and Museum the same color?"

Answer:

☐ Can't answer this question

Text Facts

Fact 1

Thompson joined the Communist Party in 1935 but left the group by 1938. In the early stages of World War II, Thompson worked at an aircraft factory. He was investigated by the FBI because of his early Communist Party affiliation. These events were fodder for his semi-autobiographical debut novel Now and on Earth (1942).

Jim Thompson (writer) - Wikipedia

Fact 2

Unsourced material may be challenged and removed. James Myers Thompson (September 27, 1906 – April 7, 1977) was an American author and screenwriter, known for his hardboiled crime fiction . Thompson wrote more than thirty novels, the majority of which were original paperback publications, published from the late-1940s through mid-1950s.

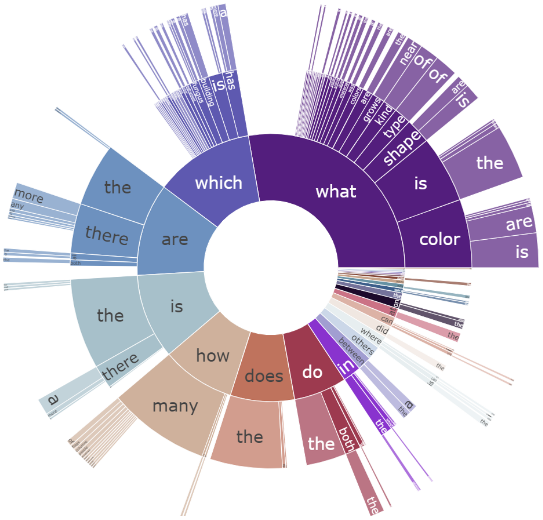
Jim Thompson (writer) - Wikipedia

Question What jobs did James Myers Thompson do during his life?

Answer:

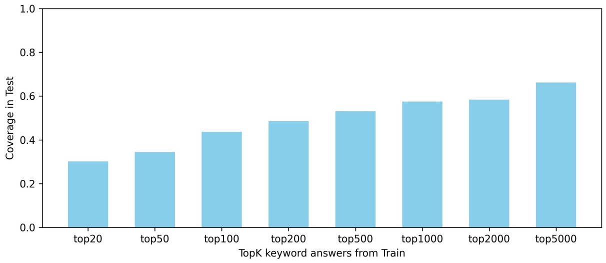
☐ Can't answer this question due to missing facts

B. Visualization of Image Question Prefixes



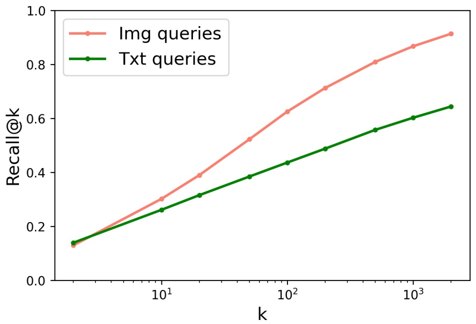
C. Classification Based Coverage

The figure below shows the test set coverage of Top-K training **keywords** (image-based). All keywords (>5k) provides only ∼70% coverage. The **full sentence** answers are almost entirely unique, suggesting that classification-based approaches are at a significant disadvantage on WebQA.



D. Additional Results on Full-scale Retrieval

Assuming known answer modality, CLIP [3] achieves 91% and 64% recall rate for image- and text-based queries when 2,000 candidates are retrieved. Without the modality knowledge, the recall rate for image-based queries is zero because the question-image similarities are systematically lower than question-text similarities. Future work may fine-tune dense multimodal retrieval models to close the gap between question-image and question-text similarities.

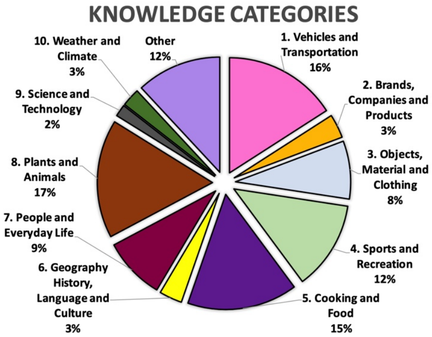


E. Comparing WebQA and recent benchmarks

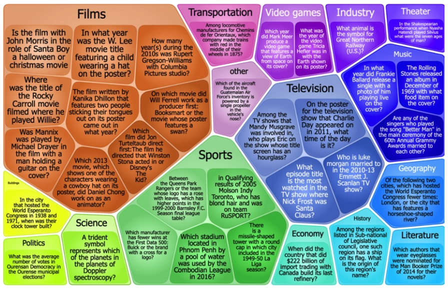
We succinctly contrast WebQA against existing knowledge-aware and multimodal datasets in the main paper. We provide here a more complete clarification of the new contributions of WebQA over relevant datasets in prior work in terms of data size, modalities and reasoning levels.

WebQA differs from QAngaroo, HotpotQA, ComplexWebQuestions, HybridQA and NaturalQuestions either in the knowledge-awareness or the involvement of both text and image modalities. OK-VQA, MultiModalQA, ManyModalQA and MIMOQA qualify as both knowledge-seeking and multimodal. Thus we explain them in detail.

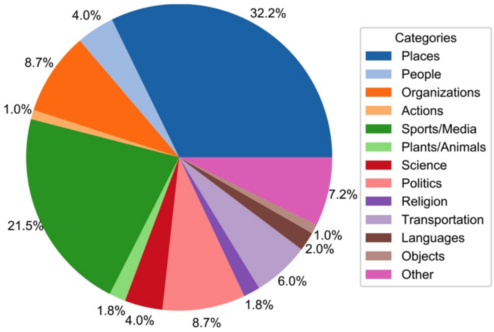
OK-VQA [2] OK-VQA and our task differ in the role of images. Images in OK-VQA are regarded as part of the query rather than the knowledge source, so source retrieval is not required. However, images in WebQA serve as the knowledge rather than part of the query and can only be processed after retrieval. OK-VQA Topics:



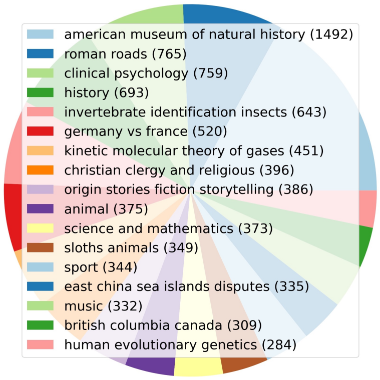
MultiModalQA [5] MultiModalQA and WebQA differ in the way qa-pairs were constructed and the answer schema. First, MultiModalQA questions are generated from templates. While this facilitates the data generation process, it does not mirror the way real users construct queries. Once the question template is detected, the task reduces to filling in blanks with modality-specific answering mechanisms. This problem-solving manner might not generalize to queries issued by real users where an underlying template is less obvious. In contrast, queries in WebQA are written by annotators, and more structurally diverse. Second, MultiModalQA requires different answer schemas for TextQA, ImageQA and TableQA. TextQA expects a span, "yes" or "no" as an answer. ImageQA expects selection from a fixed answer vocabulary determined by the training set. TableQA expects "yes", "no", a table cell, or a summary of more than one table cells via a predicted aggregation operation (i.e. SUM / MEAN / COUNT). We unify the answer schema to be a complete natural language sentence and use an open answer set, so neither span prediction nor classification over a fixed vocabulary suffice. MultiModalQA Topics:



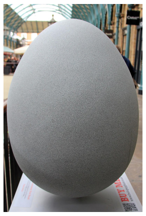
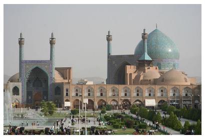
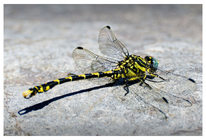
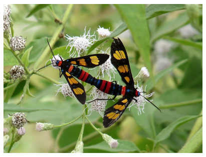
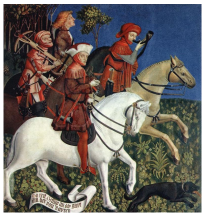
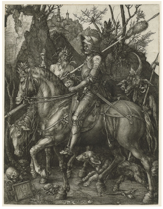
ManyModalQA [1] The primary challenge ManyModalQA addresses is the choice of answer modality – rather than knowledge aggregation or extraction. Our focus is less about distinguishing the answer modality, than about representing world knowledge in a unified space, since mastering the latter may naturally eliminate the need to classify questions into different buckets according to their answer modality. Also, to avoid ambiguity and for easy evaluation, ManyModalQA restricts all answers to be a single word. Therefore, the following question answering is a multiple choice task from [all words in the given context + a predefined answer vocabulary]. We argue that multiple choice is an unnatural simplification, because the finite and static answer space imposes a hard limit on the capacity of an answering system, especially when we consider unfamiliar domains, constant shift of world states, and unlimited coverage of the Web. This leads to us formulating WebQA as a free-form generation task, which, although it introduces new challenges for evaluation, better resembles real-world use cases and suits the needs of downstream applications such as voice assistants or conversational agents. Last but not least, ManyModelQA is much smaller than WebQA in size. ManyModalQA Topics:



MIMOQA [4] requires selecting a text span from a given context and an image from a set of related images as a multimodal output pair. However, this task formulation does not support queries whose answers should be a digested and summarized version of the given sources instead of a span. WebQA requires further information aggregation and summarization through either numerical or logical reasoning, highlighting the major advantage over MIMOQA in reasoning levels. Plus, WebQA tests natural language generation ability while MIMOQA only requires span prediction and retrieval, both under the classification banner.



F. Additional Qualitative Analysis

Source(s)	Question (**Q**), Answer Prediction (**Pred**), & Keywords (**KW**)
	**Q:** Is the surface of the egg next to the handrail at the Big Egg Hunt in Covent Garden London shiny or dull? **Pred:** The surface of the egg next to the handrail at the Big Egg Hunt in Covent Garden London is shiny. **KW:** Dull **Notes:** The model does not have a reasonably large vocabulary for visual properties. It could be the case that shiny is preferred since it is a more common word in the training set.
	**Q:** What is the color of the dome of the Isfahan Royal Mosque ? **Pred:** The color of the dome of the Isfahan Royal Mosque is white . **KW:** Blue **Notes:** It can be an issue of either not looking at the correct region, or not associating the color words with their visual appearances. But regardless, the superficial pattern matching skills do not work on the adversarial testing samples.
 	**Q:**What part of the euchromia polymena has the same coloring as the abdomen of the tiger dragonfly ' s abdomen ? **Pred:** The euchromia polymena has the same coloring as the abdomen of the tiger dragonfly ' s abdomen . **KW:** Wings **Notes:** The model does not understand the question and is treating it as binary.
 	**Q:** What animal is underneath the horses legs in both the paintings Knight , Death , and the Devil and Prince Tassilo Rides to Hunting ? **Pred:** A dog is underneath the horses legs in both the Knight , Death , and the Devil and Prince Tassilo Rides to Hunting. **KW:** A dog **Notes:** The model predicts correctly, probably due to precise object detection.
1. Coinage was used in the Ptolemaic Kingdom during the last dynasty of Egypt and, briefly, during Roman rule of Egypt . Ptolemaic coinage was struck in Phoenician weight, also known as Ptolemaic weight (about 14.2 grams). This standard, which was not used elsewhere in the Hellenistic world, was smaller than the dominant Attic weight. 2. All the male rulers of the dynasty took the name Ptolemy, while queens regnant were all called Cleopatra, Arsinoe or Berenice. The most famous member of the line was the last queen, Cleopatra VII, known for her role in the Roman political battles between Julius Caesar and Pompey, and later between Octavian and Mark Antony. **Notes:** The model picks up the correct entity	**Q:** What type of currency was used during Cleopatra VII ' s reign ? **Pred:** Ptolemaic coinage . **KW:** Ptolemaic coinage

G. Datasheet for WEBQA

G.1. Motivation

For what purpose was the dataset created?.

WEBQA was created to drive the research progress in multihop, multimodal question answering, which would bridge the gap between the natural language and vision community.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?.

The initial version of WEBQA was created by Yingshan Chang and Yonatan Bisk on behalf of Language Technology Institute, Carnegie Mellon University, and Mridu Narang at Microsoft Bing.

Who funded the creation of the dataset? . Microsoft Research and Bing provided the funds for crowdsourcing and web crawling.

G.2. Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?. Each instance is a tuple of (Knowledge Sources, Question, Answer), where a knowledge source can be either an image assisted by a caption, or a snippet. Questions and Answers are in textual form.

How many instances are there in total (of each type, if appropriate)?. WEBQA is structured as having answers that can be found either via image search or general web (text) search. So there are two folds of data, containing 22,423 image-based queries and 24,343 text-based queries, respectively. There are 600K images crawled from Wikipedia and 750K snippets crawled from the general Web (mostly from Wikipedia) serving as potential knowledge sources.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?. WEBQA is a sample of instances. It is presumably intended to be a random sample of instances representing what one might encounter during a real web search experience. Manual efforts were put in to ensure reasonable coverage and diversity. Only qualitative tests were run to show the inclusiveness.

What data does each instance consist of?. Each data instance consists of text and images.

Is there a label or target associated with each instance?. The answer component is regarded as the target. Each instance is associated with one human-written answer in the format of a complete natural language sentence. Additionally, each instance in the testing set has multiple (3-6) full sentence answers as well as a keyword answer annotated by humans, which is supposed to be a succinct rephrasing of the corresponding long-form answer.

Is any information missing from individual instances?. Everything is included.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?. There are no relationships between instances except for the fact that multiple instances may share knowledge sources.

Are there recommended data splits (e.g., training, development/validation, testing)?. The dataset comes with specified train/dev/test splits. The split on the text-based fold was determined randomly while the test split on the image-based fold was adversarialy selected to prevent spurious shortcut learning from inflating the metrics.

Are there any errors, sources of noise, or redundancies in the dataset?. Erroneous instances were pruned during the validation process after the initial collection, where we had human annotators report mistakes and inconsistency. The released version is clean.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?. No. All the information crawled from the Web was downloaded and fixed when the dataset was constructed.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)?. No. All data was derived from crowdsourcing and publicly available content on the web.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?. No, data was specifically pulled from known vetted resources (e.g. Wikipedia / Wikimedia).

Does the dataset relate to people?. No

G.3. Collection Process

How was the data associated with each instance acquired?. The questions and answers were curated by crowdsourcing. The knowledge sources were mined from the web that were directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?. Crowdsourcing relied on the Amazon Mechanical Turk platform. Web crawling was assisted by Bing Visual Search and Wikipedia APIs.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?. All question-answer pairs were human-curated. Knowledge sources for each sample are determined by their relevance to the question-answer pair.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?. Crowdworkers are paid with an average hourly wage above $13.

Over what timeframe was the data collected?. WebQA was collected and validated from Oct 2020 to Aug 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)?. No

Does the dataset relate to people?. No

G.4. Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?. After the initial collection, each sample was validated by 2 or 3 crowdworkers. Problematic samples were discarded. Testing samples with low human agreements were discarded. Besides, each sample in the image-based fold was assigned a question category label produced by a text analysis algorithm.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?. The raw unprocessed data (consisting of crowdsourcing output, history versions of unpruned dataset) is saved.

Is the software used to preprocess/clean/label the instances available?. While a script running a sequence of commands is not available, all codes used to process the data is open source on Github.

G.5. Uses

Has the dataset been used for any tasks already?. The dataset was introduced in the paper WebQA: Multihop and Multimodal QA.

Is there a repository that links to any or all papers or systems that use the dataset?. Papers using this dataset will be listed in https://webqna.github.io/ or linked from the EvalAI leaderboard.

What (other) tasks could the dataset be used for?. WebQA can be used for modelling works in the areas of knowledge retrieval, multimodal reasoning and open-domain question answering.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?. No. There is minimal known risks for harm.

Are there tasks for which the dataset should not be used?. Not to our knowledge

G.6. Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?. Yes. WebQA will be made publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?. See https://webqna.github.io/ for downloading instructions.

When will the dataset be distributed?. WebQA will be released to the public in Sep 2021.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?. The crawled data copyright belongs to the websites that the data originally appeared in (e.g. Wikimedia Foundation). WebQA will be distributed under freely to academic researchers upon request.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?. No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?. No

G.7. Maintenance

Who is supporting/hosting/maintaining the dataset?. WebQA is supported and maintained by Language Technologies Institute @CMU and Microsoft Research, and the leaderboard is hosted on EvalAI.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?. {yingshac, ybisk}@cs.cmu.edu

Is there an erratum?. All changes to the dataset will be announced on https://webqna.github.io/

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?. All updates (if necessary) will be posted on https://webqna.github.io/

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?. WebQA is not related to people.

Will older versions of the dataset continue to be supported/hosted/maintained?. All changes to the dataset will be announced on https://webqna.github.io/. Outdated versions will be kept around for consistency.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?. Any extension/augmentation by an external party should be made after contacting the original authors.

References

[1] Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886, 2020. 3

[2] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. 2

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[4] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, 2021. 3

[5] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. 3