CoQA: A Conversational Question Answering Challenge

Siva Reddy* **Danqi Chen*** **Christopher D. Manning**
Computer Science Department
Stanford University
{sivar,danqi,manning}@cs.stanford.edu

Abstract

Humans gather information through conversations involving a series of interconnected questions and answers. For machines to assist in information gathering, it is therefore essential to enable them to answer conversational questions. We introduce CoQA, a novel dataset for building **Co**nversational **Q**uestion **A**nswering systems.[1] Our dataset contains 127k questions with answers, obtained from 8k conversations about text passages from seven diverse domains. The questions are conversational, and the answers are free-form text with their corresponding evidence highlighted in the passage. We analyze CoQA in depth and show that conversational questions have challenging phenomena not present in existing reading comprehension datasets, e.g., coreference and pragmatic reasoning. We evaluate strong dialogue and reading comprehension models on CoQA. The best system obtains an F1 score of 65.4%, which is 23.4 points behind human performance (88.8%), indicating there is ample room for improvement. We present CoQA as a challenge to the community at https://stanfordnlp.github.io/coqa.

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Q_i), an answer (A_i) and a rationale (R_i) that supports the answer.

1 Introduction

We ask other people a question to either seek or test their knowledge about a subject. Depending on their answer, we follow up with another question and their second answer builds on what has already been discussed. This incremental aspect makes human conversations succinct. An inability to build and maintain common ground in this way is part of why virtual assistants usually don't seem like competent conversational partners. In this paper, we introduce CoQA, a **Co**nversational **Q**uestion **A**nswering dataset for measuring the ability of machines to participate in a question-answering style conversation. In CoQA, a machine has to understand a text passage and answer a series of questions that appear in a conversation. We develop CoQA with three main goals in mind.

The first concerns the nature of questions in a human conversation. Figure 1 shows a conversation between two humans who are reading a passage, one acting as a questioner and the other as an answerer. In this conversation, every question after the first is dependent on the conversation history.

*The first two authors contributed equally.
[1]CoQA is pronounced as *coca*.

Dataset	Conversational	Answer Type	Domain
MCTest (Richardson et al., 2013)	✗	Multiple choice	Children's stories
CNN/Daily Mail (Hermann et al., 2015)	✗	Spans	News
Children's book test (Hill et al., 2016)	✗	Multiple choice	Children's stories
SQuAD (Rajpurkar et al., 2016)	✗	Spans	Wikipedia
MS MARCO (Nguyen et al., 2016)	✗	Free-form text, Unanswerable	Web Search
NewsQA (Trischler et al., 2017)	✗	Spans	News
SearchQA (Dunn et al., 2017)	✗	Spans	Jeopardy
TriviaQA (Joshi et al., 2017)	✗	Spans	Trivia
RACE (Lai et al., 2017)	✗	Multiple choice	Mid/High School Exams
Narrative QA (Kočiskỳ et al., 2018)	✗	Free-form text	Movie Scripts, Literature
SQuAD 2.0 (Rajpurkar et al., 2018)	✗	Spans, Unanswerable	Wikipedia
CoQA (this work)	✓	Free-form text, Unanswerable; Each answer comes with a text span rationale	Children's Stories, Literature, Mid/High School Exams, News, Wikipedia, Reddit, Science

Table 1: Comparison of CoQA with existing reading comprehension datasets.

For instance, Q_5 (*Who?*) is only a single word and is impossible to answer without knowing what has already been said. Posing short questions is an effective human conversation strategy, but such questions are really difficult for machines to parse. As is well known, state-of-the-art models rely heavily on lexical similarity between a question and a passage (Chen et al., 2016; Weissenborn et al., 2017). At present, there are no large-scale reading comprehension datasets which contain questions that depend on a conversation history (see Table 1) and this is what CoQA is mainly developed for.[2]

The second goal of CoQA is to ensure the naturalness of answers in a conversation. Many existing QA datasets restrict answers to contiguous text spans in a given passage (Table 1). Such answers are not always natural, for example, there is no span-based answer to Q_4 (*How many?*) in Figure 1. In CoQA, we propose that the answers can be free-form text, while for each answer, we also provide a text span from the passage as a rationale to the answer. Therefore, the answer to Q_4 is simply *Three* while its rationale spans across multiple sentences. Free-form answers have been studied in previous reading comprehension datasets e.g., MS MARCO (Nguyen et al., 2016) and NarrativeQA (Kočiskỳ et al., 2018) and metrics such as BLEU or ROUGE are used for evaluation due to the high variance of possible answers. One key difference in our setting is that we require answerers to first select a text span as the rationale and then edit it to obtain a free-form answer.[3] Our method strikes a balance between naturalness of answers and reliable automatic evaluation, and it results in a high human agreement (88.8% F1 word overlap among human annotators).

The third goal of CoQA is to enable building QA systems that perform robustly across domains. The current QA datasets mainly focus on a single domain which makes it hard to test the generalization ability of existing models. Hence we collect our dataset from seven different domains — children's stories, literature, middle and high school English exams, news, Wikipedia, Reddit and science. The last two are used for out-of-domain evaluation.

To summarize, CoQA has the following key characteristics:

- It consists of 127k conversation turns collected from 8k conversations over text passages. The average conversation length is 15 turns, and each turn consists of a question and an answer.
- It contains free-form answers and each answer has a span-based rationale highlighted in the passage.
- Its text passages are collected from seven diverse domains: five are used for in-domain evaluation and two are used for out-of-domain evaluation.

Almost half of CoQA questions refer back to conversational history using anaphors, and a large

[2] Concurrent with our work, Choi et al. (2018) also created a conversational dataset with a similar goal, but it differs in many aspects. We discuss the details in Section 7.

[3] In contrast, in NarrativeQA, the annotators were encouraged to use their own words and copying was not allowed in their interface.

portion require pragmatic reasoning making it challenging for models that rely on lexical cues alone. We benchmark several deep neural network models, building on top of state-of-the-art conversational and reading comprehension models (Section 5). The best-performing system achieves an F1 score of 65.4%. In contrast, humans achieve 88.8% F1, 23.4% F1 higher, indicating that there is a lot of headroom for improvement.

2 Task Definition

Given a passage and a conversation so far, the task is to answer the next question in the conversation. Each turn in the conversation contains a question and an answer.

For the example in Figure 2, the conversation begins with question Q_1. We answer Q_1 with A_1 based on the evidence R_1, which is a contiguous text span from the passage. In this example, the answerer only wrote the *Governor* as the answer but selected a longer rationale *The Virginia governor's race*.

When we come to Q_2 (*Where?*), we must refer back to the conversation history otherwise its answer could be *Virginia* or *Richmond* or something else. In our task, conversation history is indispensable for answering many questions. We use conversation history Q_1 and A_1 to answer Q_2 with A_2 based on the evidence R_2. Formally, to answer Q_n, it depends on the conversation history: Q_1, A_1, ..., Q_{n-1}, A_{n-1}. For an unanswerable question, we give *unknown* as the final answer and do not highlight any rationale.

In this example, we observe that the entity of focus changes as the conversation progresses. The questioner uses *his* to refer to *Terry* in Q_4 and *he* to *Ken* in Q_5. If these are not resolved correctly, we end up with incorrect answers. The conversational nature of questions requires us to reason from multiple sentences (the current question and the previous questions or answers, and sentences from the passage). It is common that a single question may require a rationale spanning across multiple sentences (e.g., Q_1 Q_4 and Q_5 in Figure 1). We describe additional question and answer types in Section 4.

Note that we collect rationales as (optional) evidence to help answer questions. However, they are not provided at testing time. A model needs to decide on the evidence by itself and derive the final answer.

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q_1: What are the candidates **running** for?
A_1: Governor
R_1: The Virginia governor's race

Q_2: **Where**?
A_2: Virginia
R_2: The Virginia governor's race

Q_3: Who is the democratic candidate?
A_3: **Terry McAuliffe**
R_3: Democrat Terry McAuliffe

Q_4: Who is **his** opponent?
A_4: **Ken Cuccinelli**
R_4 Republican Ken Cuccinelli

Q_5: What party does **he** belong to?
A_5: Republican
R_5: Republican Ken Cuccinelli

Q_6: Which of **them** is winning?
A_6: Terry McAuliffe
R_6: Democrat Terry McAuliffe, the longtime political fixer and moneyman, hasn't trailed in a poll since May

Figure 2: A conversation showing coreference chains in color. The entity of focus changes in Q4, Q5, Q6.

3 Dataset Collection

For each conversation, we employ two annotators, a questioner and an answerer. This setup has several advantages over using a single annotator to act both as a questioner and an answerer: 1) when two annotators chat about a passage, their dialogue flow is natural; 2) when one annotator responds with a vague question or an incorrect answer, the other can raise a flag which we use to identify bad workers; and 3) the two annotators can discuss guidelines (through a separate chat window) when they have disagreements. These measures help to prevent spam and to obtain high agreement data.[4] We use Amazon Mechanical Turk (AMT) to pair workers on a passage through the ParlAI MTurk API (Miller et al., 2017).

[4]Due to AMT terms of service, we allowed a single worker to act both as a questioner and an answerer after a minute of waiting. This constitutes around 12% of the data. We include this data in the training set only.

3.1 Collection Interface

We have different interfaces for a questioner and an answerer (see Appendix). A questioner's role is to ask questions, and an answerer's role is to answer questions in addition to highlighting rationales. Both questioner and answerer sees the conversation that happened until now, i.e., questions and answers from previous turns and rationales are kept hidden. While framing a new question, we want questioners to avoid using exact words in the passage in order to increase lexical diversity. When they type a word that is already present in the passage, we alert them to paraphrase the question if possible. While answering, we want answerers to stick to the vocabulary in the passage in order to limit the number of possible answers. We encourage this by asking them to first highlight a rationale (text span), which is then automatically copied into the answer box, and we further ask them to edit the copied text to generate a natural answer. We found 78% of the answers have at least one edit such as changing a word's case or adding a punctuation.

3.2 Passage Selection

We select passages from seven diverse domains: children's stories from MCTest (Richardson et al., 2013), literature from Project Gutenberg[5], middle and high school English exams from RACE (Lai et al., 2017), news articles from CNN (Hermann et al., 2015), articles from Wikipedia, Reddit articles from the Writing Prompts dataset (Fan et al., 2018) and science articles from AI2 Science Questions (Welbl et al., 2017).

Not all passages in these domains are equally good for generating interesting conversations. A passage with just one entity often results in questions that entirely focus on that entity. Therefore, we select passages with multiple entities, events and pronominal references using Stanford CoreNLP (Manning et al., 2014). We truncate long articles to the first few paragraphs that result in around 200 words.

Table 2 shows the distribution of domains. We reserve the Reddit and Science domains for out-of-domain evaluation. For each in-domain dataset, we split the data such that there are 100 passages in the development set, 100 passages in the test set, and the rest in the training set. For each out-of-domain dataset, we only have 100 passages in the test set.

[5]Project Gutenberg https://www.gutenberg.org

Domain	#Passages	#Q/A pairs	Passage length	#Turns per passage
In-domain				
Children's Sto.	750	10.5k	211	14.0
Literature	1,815	25.5k	284	15.6
Mid/High Sch.	1,911	28.6k	306	15.0
News	1,902	28.7k	268	15.1
Wikipedia	1,821	28.0k	245	15.4
Out-of-domain				
Reddit	100	1.7k	361	16.6
Science	100	1.5k	251	15.3
Total	8,399	127k	271	15.2

Table 2: Distribution of domains in CoQA.

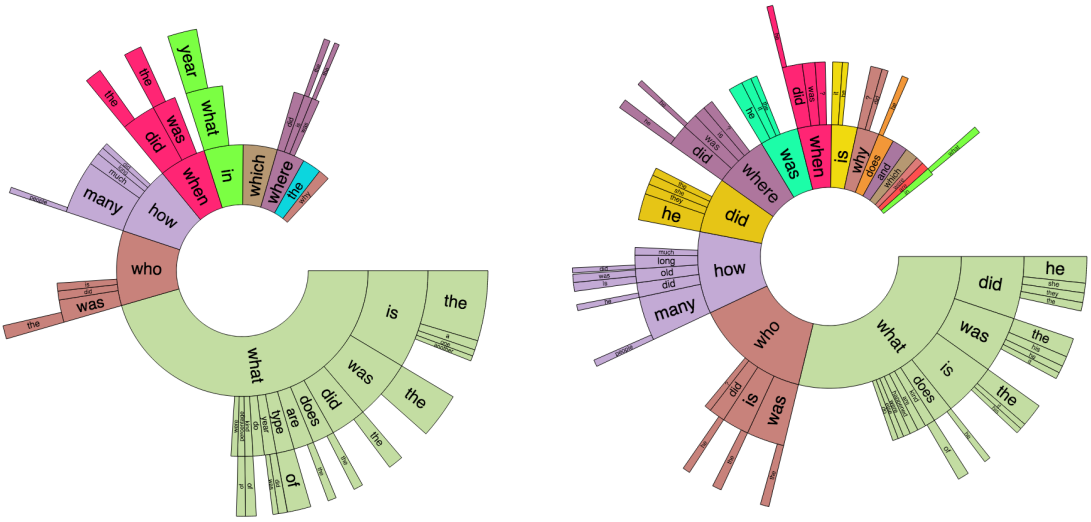
3.3 Collecting Multiple Answers

Some questions in CoQA may have multiple valid answers. For example, another answer to Q_4 in Figure 2 is *A Republican candidate*. In order to account for answer variations, we collect three additional answers for all questions in the development and test data. Since our data is conversational, questions influence answers which in turn influence the follow-up questions. In the previous example, if the original answer was *A Republican Candidate*, then the following question *Which party does he belong to?* would not have occurred in the first place. When we show questions from an existing conversation to new answerers, it is likely they will deviate from the original answers which makes the conversation incoherent. It is thus important to bring them to a common ground with the original answer.

We achieve this by turning the answer collection task into a game of predicting original answers. First, we show a question to an answerer, and when she answers it, we show the original answer and ask her to verify if her answer matches the original. For the next question, we ask her to guess the original answer and verify again. We repeat this process with the same answerer until the conversation is complete. The entire conversation history is shown at each turn (question, answer, original answer for all previous turns but not the rationales). In our pilot experiment, the human F1 score is increased by 5.4% when we use this verification setup.

4 Dataset Analysis

What makes the CoQA dataset conversational compared to existing reading comprehension datasets like SQuAD? What linguistic phenomena do the



(a) SQuAD (b) CoQA

Figure 3: Distribution of trigram prefixes of questions in SQuAD and CoQA.

questions in CoQA exhibit? How does the conversation flow from one turn to the next? We answer these questions below.

4.1 Comparison with SQuAD 2.0

SQuAD has been the main benchmark for reading comprehension. In the following, we perform an in-depth comparison of CoQA and the latest version of SQuAD (Rajpurkar et al., 2018). Figure 3(a) and Figure 3(b) show the distribution of frequent trigram prefixes. Because of the free-form nature of answers, we expect a richer variety of questions in CoQA than that in SQuAD. While nearly half of SQuAD questions are dominated by *what* questions, the distribution of CoQA is spread across multiple question types. Several sectors indicated by prefixes *did, was, is, does* and *and* are frequent in CoQA but are completely absent in SQuAD. While coreferences are non-existent in SQuAD, almost every sector of CoQA contains coreferences (*he, him, she, it, they*) indicating CoQA is highly conversational.

Since a conversation is spread over multiple turns, we expect conversational questions and answers to be shorter than in a standalone interaction. In fact, questions in CoQA can be made up of just one or two words (*who?*, *when?*, *why?*). As seen in Table 3, on average, a question in CoQA is only 5.5 words long while it is 10.1 for SQuAD. The answers are a bit shorter in CoQA than SQuAD because of the free-form nature of the answers.

Table 4 provides insights into the type of answers in SQuAD and CoQA. While the original version of SQuAD (Rajpurkar et al., 2016) does not have any unanswerable questions, the later version (Rajpurkar et al., 2018) focuses solely on obtaining them resulting in higher frequency than in CoQA. SQuAD has 100% span-based answers by design, whereas in CoQA, 66.8% of the answers overlap with the passage after ignoring punctuation and case mismatches.[6] The rest of the answers, 33.2%, do not exactly overlap with the passage (see Section 4.3). It is worth noting that CoQA has 11.1% and 8.7% questions with *yes* or *no* as answers whereas SQuAD has 0%. Both datasets have a high number of named entities and noun phrases as answers.

4.2 Linguistic Phenomena

We further analyze the questions for their relationship with the passages and the conversation history. We sample 150 questions in the development set and annotate various phenomena as shown in Table 5.

If a question contains at least one content word that appears in the rationale, we classify it as *lexical match*. These comprise around 29.8% of the questions. If it has no lexical match but is a paraphrase of the rationale, we classify it as *paraphrasing*. These questions contain phenomena such as synonymy, antonymy, hypernymy, hyponymy and negation. These constitute a large portion of ques-

[6] If punctuation and case are not ignored, only 37% of the answers can be found as spans.

	SQuAD	CoQA
Passage Length	117	271
Question Length	10.1	5.5
Answer Length	3.2	2.7

Table 3: Average number of words in passage, question and answer in SQuAD and CoQA.

	SQuAD	CoQA
Answerable	66.7%	98.7%
Unanswerable	33.3%	1.3%
Span found	100.0%	66.8%
No span found	0.0%	33.2%
Named Entity	35.9%	28.7%
Noun Phrase	25.0%	19.6%
Yes	0.0%	11.1%
No	0.1%	8.7%
Number	16.5%	9.8%
Date/Time	7.1%	3.9%
Other	15.5%	18.1%

Table 4: Distribution of answer types in SQuAD and CoQA.

Phenomenon	Example	Percentage
Relationship between a question and its passage		
Lexical match	Q: Who had to rescue her? A: the coast guard R: Outen was rescued by the coast guard	29.8%
Paraphrasing	Q: Did the wild dog **approach**? A: Yes R: he **drew** cautiously **closer**	43.0%
Pragmatics	Q: Is Joey a male or female? A: Male R: it looked like a stick man so she kept **him**. She named her new noodle friend Joey	27.2%
Relationship between a question and its conversation history		
No coref.	Q: What is IFL?	30.5%
Explicit coref.	Q: Who had Bashti forgotten? A: the puppy Q: What was **his** name?	49.7%
Implicit coref.	Q: When will Sirisena be sworn in? A: 6 p.m local time Q: **Where**?	19.8%

Table 5: Linguistic phenomena in CoQA questions.

tions, around 43.0%. The rest, 27.2%, have no lexical cues, and we classify them as *pragmatics*. These include phenomena like common sense and presupposition. For example, the question *Was he loud and boisterous?* is not a direct paraphrase of the rationale *he dropped his feet with the lithe softness of a cat* but the rationale combined with world knowledge can answer this question.

For the relationship between a question and its conversation history, we classify questions into whether they are dependent or independent on the conversation history. If dependent, whether the questions contain an explicit marker or not. Our analysis shows that around 30.5% questions do not rely on coreference with the conversational history and are answerable on their own. Almost half of the questions (49.7%) contain explicit coreference markers such as *he, she, it*. These either refer to an entity or an event introduced in the conversation. The remaining 19.8% do not have explicit coreference markers but refer to an entity or event implicitly (these are often cases of *ellipsis*, as in the examples in Table 5).

4.3 Analysis of Free-form Answers

Due to the free-form nature of CoQA's answers, around 33.2% of them do not exactly overlap with the given passage. We analyze 100 conversations to study the behavior of such answers.[7] As shown in Table 6, the answers *Yes* and *No* constitute 48.5% and 30.3% respectively, totaling 78.8%. The next majority, around 14.3%, are edits to text spans to improve the fluency (naturalness) of answers. More than two thirds of these edits are just one word edits, either inserting or deleting a word. This indicates that text spans are a good approximation for natural answers, positive news for span-based reading comprehension models. The remaining one third involve multiple edits. Although multiple edits are challenging to evaluate using automatic metrics, we observe that many of these answers partially overlap with passage, indicating that word overlap is still a reliable automatic evaluation metric in our setting. The rest of the answers include counting (5.1%) and selecting a choice from the question (1.8%).

4.4 Conversation Flow

A coherent conversation must have smooth transitions between turns. We expect the narrative structure of the passage to influence our conversation flow. We split each passage into 10 uniform chunks, and identify chunks of interest in a given turn and its transition based on rationale spans. Figure 4

[7] We only pick the questions in which none of its answers can be found as a span in the passage.

Answer Type	Example	Percentage
Yes	Q: is MedlinePlus optimized for mobile? A: Yes R: There is also a site optimized for display on mobile devices	48.5%
No	Q: Is it played outside? A: No R: AFL is the highest level of professional indoor American football	30.3%
Fluency	Q: Why? A: so the investigation could continue R: while the investigation continued	14.3%
Counting	Q: how many languages is it offered in? A: Two R: The service provides curated consumer health information in English and Spanish	5.1%
Multiple choice	Q: Is Jenny older or younger? A: Older R: her baby sister is crying so loud that Jenny can't hear herself	1.8%
	Fine grained breakdown of Fluency	
Multiple edits	Q: What did she try just before that? A: She **gave** her **a** toy **horse**. R: She **would give** her **baby sister one of her** toy horses. (morphology: give → gave, horses → horse; delete: would, baby sister one of her; insert: a)	41.4%
Coreference insertion	Q: what is the cost to end users? A: **It** is free R: The service is funded by the NLM and **is free** to users	16.0%
Morphology	Q: Who was messing up the neighborhoods? A: vandals R: **vandalism** in the neighborhoods	13.9%
Article insertion	Q: What would they cut with? A: **an** ax R: the heavy ax	7.2%
Adverb insertion	Q: How old was the diary? A: 190 years **old** R: kept 190 years ago	4.2%
Adjective deletion	Q: What type of book? A: A diary. R: a **120-page** diary	4.2%
Preposition insertion	how long did it take to get to the fire? A: **Until** supper time! R: By the time they arrived, it was almost supper time.	3.4%
Adverb deletion	Q: What had happened to the ice? A: It had changed R: It had **somewhat** changed its formation when they approached it	3.0%
Conjunction insertion	Q: what else do they get for their work? A: potatoes **and** carrots R: paid well, both in potatoes, carrots	1.3%
Noun insertion	Q: Who did A: Comedy Central **employee** R: But it was a Comedy Central account	1.3%
Coreference deletion	Q: What is the story about? A: A girl and a dog R: This is the story of a young girl and **her** dog	1.2%
Noun deletion	Q: What is the ranking in the country in terms of people studying? A: the fourth largest population R: and has the fourth largest **student** population	0.8%
Possesive insertion	Q: Whose diary was it? A: Deborah Logan**'s** R: a 120-page diary kept 190 years ago by Deborah Logan	0.8%
Article deletion	Q: why? A: They were going to the circus R: They **all** were going to the circus to see the clowns	0.8%

Table 6: Analysis of answers which don't overlap with passage.

shows the conversation flow of the first 10 turns. The starting turns tend to focus on the first few chunks and as the conversation advances, the focus shifts to the later chunks. Moreover, the turn transitions are smooth, with the focus often remaining in the same chunk or moving to a neighboring chunk. Most frequent transitions happen to the first and the last chunks, and likewise these chunks have diverse outward transitions.

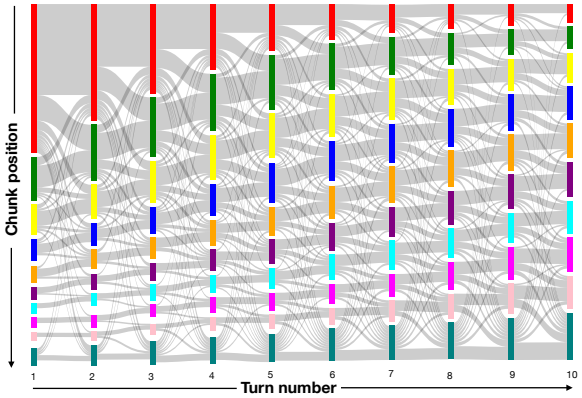


Figure 4: Chunks of interest as a conversation progresses. Each chunk is one tenth of a passage. The x-axis indicates the turn number and the y-axis indicates the chunk containing the rationale. The height of a chunk indicates the concentration of conversation in that chunk. The width of the bands is proportional to the frequency of transition between chunks from one turn to the next.

5 Models

Given a passage p, the conversation history $\{q_1, a_1, \ldots q_{i-1}, a_{i-1}\}$ and a question q_i, the task is to predict the answer a_i. Gold answers a_1, a_2, $\ldots$, a_{i-1} are used to predict a_i, similar to the setup discussed in Section 3.3.

Our task can either be modeled as a conversational response generation problem or a reading comprehension problem. We evaluate strong baselines from each modeling type and a combination of the two on CoQA.

5.1 Conversational Models

Sequence-to-sequence (*seq2seq*) models have shown promising results for generating conversational responses (Vinyals and Le, 2015; Serban et al., 2016; Zhang et al., 2018). Motivated by their success, we use a sequence-to-sequence with attention model for generating answers (Bahdanau et al., 2015). We append the conversation history and the current question to the passage, as p $<q>$ q_{i-n} $<a>$ a_{i-n} $\ldots$ $<q>$ q_{i-1} $<a>$ a_{i-1} $<q>$ q_i, and feed it into a bidirectional LSTM encoder, where n is the size of the history to be used. We generate the answer using an LSTM decoder which attends to the encoder states. Additionally, as the answer words are likely to appear in the original passage, we employ a copy mechanism in the decoder which allows to (optionally) copy a word from the passage (Gu et al., 2016; See et al., 2017). This model is referred to as the Pointer-Generator network, PGNet.

5.2 Reading Comprehension Models

The state-of-the-art reading comprehension models for extractive question answering focus on finding a span in the passage which matches the question best (Seo et al., 2016; Chen et al., 2017; Yu et al., 2018). Since their answers are limited to spans, they cannot handle questions whose answers do not overlap with the passage, e.g., Q_3, Q_4 and Q_5 in Figure 1. However this limitation makes them more effective learners than conversational models which have to generate an answer from a large space of pre-defined vocabulary.

We use the Document Reader (DrQA) model of Chen et al. (2017), which has demonstrated strong performance on multiple datasets (Rajpurkar et al., 2016; Labutov et al., 2018). Since DrQA requires text spans as answers during training, we select the span which has the highest lexical overlap (F1 score) with the original answer as the gold answer. If the answer appears multiple times in the story we use the rationale to find the correct one. If any answer word does not appear in the story, we fall back to an additional *unknown* token as the answer (about 17% in the training set). We prepend each question with its past questions and answers to account for conversation history, similar to the conversational models.

Considering that a significant portion of answers in our dataset are *yes* or *no* (Table 4), we also include an augmented reading comprehension model for comparison. We add two additional tokens, *yes* and *no*, to the end of the passage — if the gold answer is *yes* or *no*, the model is required to predict the corresponding token as the gold span; otherwise it does the same as the previous model. We refer to this model as Augmented DrQA.

	In-domain					Out-of-dom.		In-domain	Out-of-dom.	Overall
	Child.	Liter.	Mid-High.	News	Wiki.	Reddit	Science	Overall	Overall	
Development data										
Seq2seq	30.6	26.7	28.3	26.3	26.1	N/A	N/A	27.5	N/A	27.5
PGNet	49.7	42.4	44.8	45.5	45.0	N/A	N/A	45.4	N/A	45.4
DrQA	52.4	52.6	51.4	56.8	60.3	N/A	N/A	54.7	N/A	54.7
Augmt. DrQA	**67.0**	**63.2**	**63.9**	**69.8**	72.0	N/A	N/A	**67.2**	N/A	**67.2**
DrQA+PGNet	64.5	62.0	63.8	68.0	**72.6**	N/A	N/A	66.2	N/A	66.2
Human	90.7	88.3	89.1	89.9	90.9	N/A	N/A	89.8	N/A	89.8
Test data										
Seq2seq	32.8	25.6	28.0	27.0	25.3	25.6	20.1	27.7	23.0	26.3
PGNet	49.0	43.3	47.5	47.5	45.1	38.6	38.1	46.4	38.3	44.1
DrQA	46.7	53.9	54.1	57.8	59.4	45.0	51.0	54.5	47.9	52.6
Augmt. DrQA	**66.0**	63.3	66.2	**71.0**	71.3	57.7	63.0	**67.6**	60.2	**65.4**
DrQA+PGNet	64.2	**63.7**	**67.1**	68.3	**71.4**	**57.8**	**63.1**	67.0	**60.4**	65.1
Human	90.2	88.4	89.8	88.6	89.9	86.7	88.1	89.4	87.4	88.8

Table 7: Models and human performance (F1 score) on the development and the test data.

5.3 A Combined Model

Finally, we propose a model which combines the advantages from both conversational models and extractive reading comprehension models. We use DrQA with PGNet in a combined model, in which DrQA first points to the answer evidence in the text, and PGNet naturalizes the evidence into an answer. For example, for Q_5 in Figure 1, we expect that DrQA first predicts the rationale R_5, and then PGNet generates A_5 from R_5.

We make a few changes to DrQA and PGNet based on empirical performance. For DrQA, we require the model to predict the answer directly if the answer is a substring of the rationale, and to predict the rationale otherwise. For PGNet, we provide the current question and DrQA's span predictions as input to the encoder and the decoder aims to predict the final answer.[8]

6 Evaluation

6.1 Evaluation Metric

Following SQuAD, we use macro-average F1 score of word overlap as our main evaluation metric.[9] We use the gold answers of history to predict the next answer. In SQuAD, for computing a model's performance, each individual prediction is compared against n human answers resulting in n F1 scores, the maximum of which is chosen as the prediction's F1.[10] For each question, we average out F1 across these n sets, both for humans and models. In our final evaluation, we use $n = 4$ human answers for every question (the original answer and 3 additionally collected answers). The articles *a, an* and *the* and punctuations are excluded in evaluation.

6.2 Experimental Setup

For all the experiments of seq2seq and PGNet, we use the OpenNMT toolkit (Klein et al., 2017) and its default settings: 2-layers of LSTMs with 500 hidden units for both the encoder and the decoder. The models are optimized using SGD, with an initial learning rate of 1.0 and a decay rate of 0.5. A dropout rate of 0.3 is applied to all layers.

For the DrQA experiments, we use the implementation from the original paper (Chen et al., 2017). We tune the hyperparameters on the development data: the number of turns to use from the conversation history, the number of layers, number of each hidden units per layer and dropout rate. The best configuration we find is 3 layers of LSTMs with 300 hidden units for each layer. A dropout rate of 0.4 is applied to all LSTM layers and a dropout rate of 0.5 is applied to word embeddings. We used Adam to optimize DrQA models.

We initialized the word projection matrix with GloVe (Pennington et al., 2014) for conversational models and fastText (Bojanowski et al., 2017) for reading comprehension models, based on empirical performance. We update the projection matrix during training in order to learn embeddings for delimiters such as $<q>$.

[8] We feed DrQA's oracle spans into PGNet during training.

[9] SQuAD also uses exact-match metric, however we think F1 is more appropriate for our dataset because of the free-form answers.

[10] However, for computing human performance, a human prediction is only compared against $n - 1$ human answers, resulting in underestimating human performance. We fix this bias by partitioning n human answers into n different sets, each set containing $n-1$ answers, similar to Choi et al. (2018).

6.3 Results and Discussion

Table 7 presents the results of the models on the development and test data. Considering the results on the test set, the seq2seq model performs the worst, generating frequently occurring answers irrespective of whether these answers appear in the passage or not, a well known behavior of conversational models (Li et al., 2016). PGNet alleviates the frequent response problem by focusing on the vocabulary in the passage and it outperforms seq2seq by 17.8 points. However, it still lags behind DrQA by 8.5 points. A reason could be that PGNet has to memorize the whole passage before answering a question, a huge overhead which DrQA avoids. But DrQA fails miserably in answering questions with answers which do not overlap with the passage (see row *No span found* in Table 8). The augmented DrQA circumvents this problem with additional yes/no tokens, giving it a boost of 12.8 points. When DrQA is fed into PGNet, we empower both DrQA and PGNet — DrQA in producing free-form answers; PGNet in focusing on the rationale instead of the passage. This combination outperforms vanilla PGNet and DrQA models by 21.0 and 12.5 points respectively, and is competitive with the augmented DrQA (65.1 vs. 65.4).

Models vs. Humans The human performance on the test data is 88.8 F1, a strong agreement indicating that the CoQA's questions have concrete answers. Our best model is 23.4 points behind humans.

In-domain vs. Out-of-domain All models perform worse on out-of-domain datasets compared to in-domain datasets. The best model drops by 6.6 points. For in-domain results, both the best model and humans find the literature domain harder than the others since literature's vocabulary requires proficiency in English. For out-of-domain results, the Reddit domain is apparently harder. While humans achieve high performance on children's stories, models perform poorly, probably due to the fewer training examples in this domain compared to others.[11] Both humans and models find Wikipedia easy.

Error Analysis Table 8 presents fine-grained results of models and humans on the development set.

[11] We collect children's stories from MCTest which contains only 660 passages in total, of which we use 200 stories for the development and the test sets.

Type	Seq2seq	PGNet	DrQA	Augmt. DrQA	DrQA+ PGNet	Human
Answer Type						
Answerable	27.5	45.4	54.7	67.3	66.3	89.9
Unanswerable	33.9	38.2	55.0	49.1	51.2	72.3
Span found	20.2	43.6	69.8	71.0	70.5	91.1
No span found	43.1	49.0	22.7	59.4	57.0	86.8
Named Entity	21.9	43.0	72.6	73.5	72.2	92.2
Noun Phrase	17.2	37.2	64.9	65.3	64.1	88.6
Yes	69.6	69.9	7.9	75.7	72.7	95.6
No	60.2	60.3	18.4	59.6	58.7	95.7
Number	15.0	48.6	66.3	69.0	71.7	91.2
Date/Time	13.7	50.2	79.0	83.3	79.1	91.5
Other	14.1	33.7	53.5	55.6	55.2	80.8
Question Type						
Lexical Mat.	20.7	40.7	57.2	75.5	65.7	91.7
Paraphrasing	23.7	33.9	46.9	62.6	64.4	88.8
Pragmatics	33.9	43.1	57.4	64.1	60.6	84.2
No coref.	16.1	31.7	54.3	70.9	58.8	90.3
Exp. coref.	30.4	42.3	49.0	63.4	66.7	87.1
Imp. coref.	31.4	39.0	60.1	70.6	65.3	88.7

Table 8: Fine-grained results of different question and answer types in the development set. For the question type results, we only analyze 150 questions as described in Section 4.2.

We observe that humans have the highest disagreement on the unanswerable questions. The human agreement on answers which do no overlap with passage is lower than on answers which overlap. This is expected because our evaluation metric is based on word overlap rather than on the meaning of words. For the question *did Jenny like her new room?*, human answers *she loved it* and *yes* are both accepted. Finding the perfect evaluation metric for abstractive responses is still a challenging problem (Liu et al., 2016; Chaganty et al., 2018) and beyond the scope of our work. For our models' performance, seq2seq and PGNet perform well on non-overlapping answers, and DrQA performs well on overlapping answers, due to their respective designs. The augmented and combined models improve on both categories.

Among the different question types, humans find lexical matches the easiest followed by paraphrasing, and pragmatics the hardest — this is expected since questions with lexical matches and paraphrasing share some similarity with the passage, thus making them relatively easier to answer than pragmatic questions. This is also the case with the combined model, but we could not explain the be-

haviour of other models. While humans find the questions without coreferences easier than those with coreferences, the models behave sporadically. Humans find implicit coreferences easier than explicit coreferences. A conjecture is that implicit coreferences depend directly on the previous turn whereas explicit coreferences may have long distance dependency on the conversation.

Importance of conversation history Finally, we examine how important the conversation history is for the dataset. Table 9 presents the results with a varied number of previous turns used as conversation history. All models succeed at leveraging history but the gains are little beyond one previous turn. As we increase the history size, the performance decreases.

We also perform an experiment on humans to measure the trade-off between their performance and the number of previous turns shown. Based on the heuristic that short questions likely depend on the conversation history, we sample 300 one or two word questions, and collect answers to these varying the number of previous turns shown.

When we do not show any history, human performance drops to 19.9 F1 as opposed to 86.4 F1 when full history is shown. When the previous turn (question and answer) is shown, their performance boosts to 79.8 F1, suggesting that the previous turn plays an important role in understanding the current question. If the last two turns are shown, they reach up to 85.3 F1, almost close to the performance when the full history is shown. This suggests that most questions in a conversation have a limited dependency within a bound of two turns.

Augmented DrQA vs. Combined Model Although the performance of the augmented DrQA is a bit better (0.3 F1 on the testing set) than the combined model, the latter model has the following benefits: 1) The combined model provides a rationale for every answer, which can be used to justify whether the answer is correct or not (e.g., yes/no questions); and 2) we don't have to decide on the set of augmented classes beforehand which helps in answering a wide range of questions like counting and multiple choice (Table 10). We also look closer into the outputs of the two models. Although the combined model is still far from perfect, it does correctly as desired in many examples, e.g., for a counting question, it predicts a rationale *current affairs , politics , and culture* and generates an answer *three*; for a question *With who?*, it predicts a rationale *Mary and her husband , Rick* and then compresses it into *Mary and Rick* for improving the fluency; and for a multiple choice question *Does this help or hurt their memory of the event?* it predicts a rationale *this obsession may prevent their brains from remembering* and answers *hurt*. We think there is still great room for improving the combined model and we leave it to future work.

History size	Seq2seq	PGNet	DrQA	Augmt. DrQA	DrQA+ PGNet
0	24.0	41.3	50.4	62.7	61.5
1	**27.5**	43.9	**54.7**	66.8	**66.2**
2	21.4	44.6	54.6	**67.2**	66.0
all	21.0	**45.4**	52.3	64.5	64.3

Table 9: Results on the development set with different history sizes. History size indicates the number of previous turns prepended to the current question. Each turn contains a question and its answer.

	Augmt. DrQA	DrQA+ PGNet	Human
Yes	**76.2**	72.5	97.7
No	**64.0**	57.5	96.8
Fluency	**37.6**	32.3	77.2
Counting	8.8	**24.8**	88.3
Multiple choice	0.0	**46.4**	94.3

Table 10: Error analysis of questions with answers which do not overlap with the text passage.

7 Related work

We organize CoQA's relation to existing work under the following criteria.

Knowledge source We answer questions about text passages — our knowledge source. Another common knowledge source is machine-friendly databases which organize world facts in the form of a table or a graph (Berant et al., 2013; Pasupat and Liang, 2015; Bordes et al., 2015; Saha et al., 2018; Talmor and Berant, 2018). However understanding their structure requires expertise, making it challenging to crowd-source large QA datasets without relying on templates. Like passages, other human friendly sources are images and videos (Antol et al., 2015; Das et al., 2017; Hori et al., 2018).

Naturalness There are various ways to curate questions: removing words from a declarative sentence to create a fill-in-the-blank question (Her-

mann et al., 2015), using a hand-written grammar to create artificial questions (Weston et al., 2016; Welbl et al., 2018), paraphrasing artificial questions to natural questions (Saha et al., 2018; Talmor and Berant, 2018) or, in our case, letting humans ask natural questions (Rajpurkar et al., 2016; Nguyen et al., 2016). While the former enable collecting large and cheap datasets, the latter enable collecting natural questions.

Recent efforts emphasize collecting questions without seeing the knowledge source in order to encourage the independence of question and documents (Joshi et al., 2017; Dunn et al., 2017; Kočiskỳ et al., 2018). Since we allow a questioner to see the passage, we incorporate measures to increase independence, although complete independence is not attainable in our setup (Section 3.1). However, an advantage of our setup is that the questioner can validate the answerer on the spot resulting in high agreement data.

Conversational Modeling Our focus is on questions that appear in a conversation. Iyyer et al. (2017) and Talmor and Berant (2018) break down a complex question into a series of simple questions mimicking conversational QA. Our work is closest to Das et al. (2017) and Saha et al. (2018) who perform conversational QA on images and a knowledge graph respectively, with the latter focusing on questions obtained by paraphrasing templates.

In parallel to our work, Choi et al. (2018) also created a dataset of conversations in the form of questions and answers on text passages. In our interface, we show a passage to both the questioner and the answerer, whereas their interface only shows a title to the questioner and the full passage to the answerer. Since their setup encourages the answerer to reveal more information for the following questions, their average answer length is 15.1 words (our average is 2.7). While the human performance on our test set is 88.8 F1, theirs is 74.6 F1. Moreover, while CoQA's answers can be free-form text, their answers are restricted only to extractive text spans. Our dataset contains passages from seven diverse domains, whereas their dataset is built only from Wikipedia articles about people.

Concurrently, Saeidi et al. (2018) created a conversational QA dataset for regulatory text such as tax and visa regulations. Their answers are limited to *yes* or *no* along with a positive characteristic of permitting to ask clarification questions when a given question cannot be answered. Elgohary et al. (2018) proposed a sequential question answering dataset collected from Quiz Bowl tournaments, where a sequence contains multiple related questions. These questions are related to the same concept while not focusing on the dialogue aspects (e.g., coreference). Zhou et al. (2018) is another dialogue dataset based on a single movie-related Wikipedia article, in which two workers are asked to chat about the content. Their dataset is more like chit-chat style conversations while our dataset focuses on multi-turn question answering.

Reasoning Our dataset is a testbed of various reasoning phenomena occurring in the context of a conversation (Section 4). Our work parallels a growing interest in developing datasets that test specific reasoning abilities: algebraic reasoning (Clark, 2015), logical reasoning (Weston et al., 2016), common sense reasoning (Ostermann et al., 2018) and multi-fact reasoning (Welbl et al., 2018; Khashabi et al., 2018; Talmor and Berant, 2018).

Recent progress on CoQA Since we first released the dataset in August 2018, the progress of developing better models on CoQA has been rapid. Instead of simply prepending the current question with its previous questions and answers, Huang et al. (2019) proposed a more sophisticated solution to effectively stack single-turn models along the conversational flow. Others (e.g., Zhu et al., 2018) attempted to incorporate the most recent pretrained language representation model BERT (Devlin et al., 2018)[12] into CoQA and demonstrated superior results. As of the time we finalized the paper (Jan 8, 2019), the state-of-art F1 score on the test set was 82.8.

8 Conclusions

In this paper, we introduced CoQA, a large scale dataset for building conversational question answering systems. Unlike existing reading comprehension datasets, CoQA contains conversational questions, free-form answers along with text spans as rationales, and text passages from seven diverse domains. We hope this work will stir more research in conversational modeling, a key ingredient for enabling natural human-machine communication.

[12]Pretrained BERT models were released in November 2018, which have demonstrated large improvements across a wide variety of NLP tasks.

Acknowledgements

We would like to thank MTurk workers, especially the Master Chatters and the MTC forum members, for contributing to the creation of CoQA, for giving feedback on various pilot interfaces, and for promoting our hits enthusiastically on various forums. CoQA has been made possible with financial support from the Facebook ParlAI and the Amazon Research awards, and gift funding from Toyota Research Institute. Danqi is supported by a Facebook PhD fellowship. We also would like to thank the members of the Stanford NLP group for critical feedback on the interface and experiments. We especially thank Drew Arad Hudson for participating in initial discussions, and Matthew Lamm for proofreading the paper. We also thank the VQA team and Spandana Gella for their help in generating Figure 3.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*, San Diego, California.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1533–1544, Seattle, Washington.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale Simple Question Answering with Memory Networks. *arXiv preprint arXiv:1506.02075*.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Association for Computational Linguistics (ACL)*, pages 643–653, Melbourne, Australia.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. *Association for Computational Linguistics (ACL)*, pages 2358–2367.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Association for Computational Linguistics (ACL)*, pages 1870–1879, Vancouver, Canada.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2174–2184, Brussels, Belgium.

Peter Clark. 2015. Elementary School Science and Math Tests as a Driver for AI: Take the Aristo Challenge! In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 4019–4021, Austin, Texas.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Computer Vision and Pattern Recognition (CVPR)*, pages 326–335, Honolulu, Hawaii.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. Dataset and Baselines for Sequential Open-Domain Question Answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1077–1083, Brussels, Belgium.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Association for Computational Linguistics (ACL)*, pages 889–898, Melbourne, Australia.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In *Association for Computational Linguistics (ACL)*, pages 1631–1640, Berlin, Germany.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1693–1701, Montreal, Canada.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.

Chiori Hori, Huda Alamri, Jue Wang, Gordon Winchern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, and others. 2018. End-to-End Audio Visual Scene-Aware Dialog using Multimodal Attention-Based Video Features. *arXiv preprint arXiv:1806.08409*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. FlowQA: Grasping Flow in History for Conversational Machine Comprehension. In *International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based Neural Structured Learning for Sequential Question Answering. In *Association for Computational Linguistics (ACL)*, pages 1821–1831, Vancouver, Canada.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Association for Computational Linguistics (ACL)*, pages 1601–1611, Vancouver, Canada.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 252–262, New Orleans, Louisiana.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Association for Computational Linguistics (ACL)*, pages 67–72, Vancouver, Canada.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Igor Labutov, Bishan Yang, Anusha Prakash, and Amos Azaria. 2018. Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds. In *Association for Computational Linguistics (ACL)*, pages 833–844, Melbourne, Australia.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 110–119, San Diego, California.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2122–2132, Austin, Texas.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP

Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL)*, pages 55–60, Baltimore, Maryland.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A Dialog Research Software Platform. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–84, Copenhagen, Denmark.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *International Workshop on Semantic Evaluation (SemEval)*, pages 747–757, New Orleans, Louisiana.

Panupong Pasupat and Percy Liang. 2015. Compositional Semantic Parsing on Semi-Structured Tables. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1470–1480, Beijing, China.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Association for Computational Linguistics (ACL)*, pages 784–789, Melbourne, Australia.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 193–203, Seattle, Washington.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2087–2097, Brussels, Belgium.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 705–713, New Orleans, Louisiana.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083, Vancouver, Canada.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. *arXiv preprint arXiv:1611.06216*.

Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 641–651, New Orleans, Louisiana.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada.

Oriol Vinyals and Quoc Le. 2015. A Neural Conversational Model. *arXiv preprint arXiv:1506.05869*.

Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making Neural QA as Simple as Possible but not Simpler. In *Computational Natural Language Learning (CoNLL)*, pages 271–280, Vancouver, Canada.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing Multiple Choice Science Questions. In *Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Fast and Accurate Reading Comprehension by Combining Self-Attention and Convolution. In *International Conference on Learning Representations (ICLR)*, Vancouver, Canada.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *Association for Computational Linguistics (ACL)*, pages 2204–2213, Melbourne, Australia.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A Dataset for Document Grounded Conversations. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 708–713, Brussels, Belgium.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering. *arXiv preprint arXiv:1812.03593*.

Appendix

Worker Selection

First each worker has to pass a qualification test that assesses their understanding of the guidelines of conversational QA. The success rate for the qualification test is 57% with 960 attempted workers. The guidelines indicate this is a conversation about a passage in the form of questions and answers, an example conversation and do's and don'ts. However, we give complete freedom for the workers to judge what is good and bad during the real conversation. This helped us in curating diverse categories of questions that were not present in the guidelines (e.g., true or false, fill in the blank and time series questions). We pay workers an hourly wage around 8–15 USD.

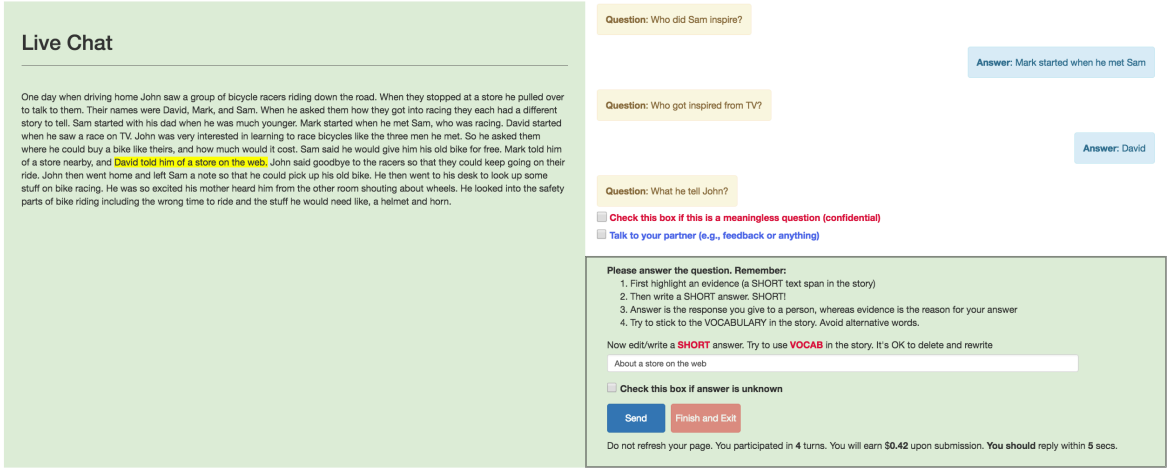
Annotation Interface

Figure 5 shows the annotation interfaces for both questioners and answerers.

Additional Examples

We provide additional examples in Figure 6 and Figure 7.

Questioner Interface



Answerer Interface

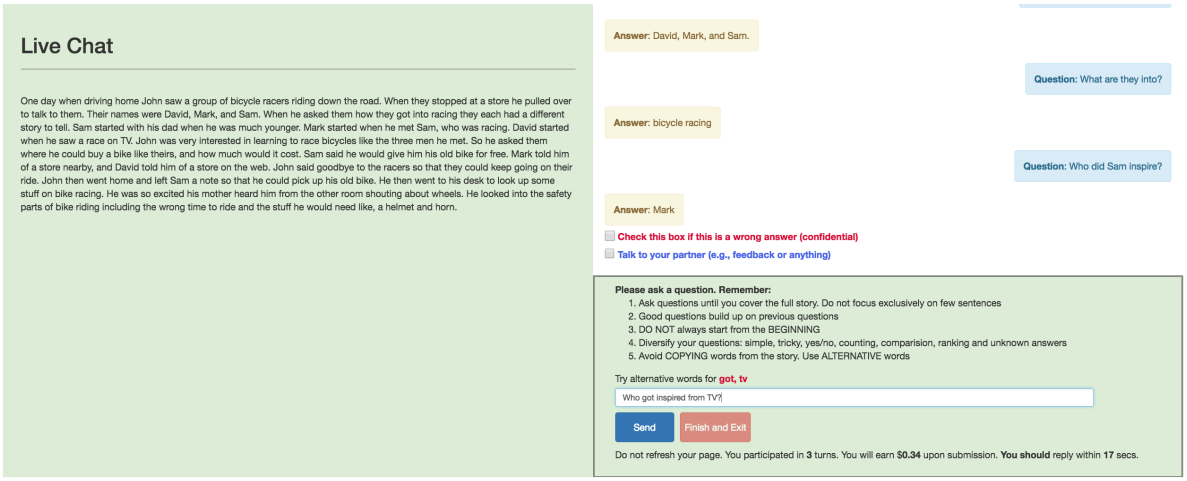


Figure 5: Annotation interfaces for questioner (top) and answerer (bottom).

Anthropology is the study of humans and their societies in the past and present. Its main subdivisions are social anthropology and cultural anthropology, which describes the workings of societies around the world, ... Similar organizations in other countries followed: The American Anthropological Association in 1902, the Anthropological Society of Madrid (1865), the Anthropological Society of Vienna (1870), the Italian Society of Anthropology and Ethnology (1871), and many others subsequently. The majority of these were evolutionist. One notable exception was the Berlin Society of Anthropology (1869) founded by Rudolph Virchow, known for his vituperative attacks on the evolutionists. Not religious himself, he insisted that Darwin's conclusions lacked empirical foundation.

Q: Who disagreed with Darwin?
A: Rudolph Virchow
R: Rudolph Virchow, known for his vituperative attacks on the evolutionists. Not religious himself, he insisted that Darwin's conclusions lacked empirical foundation.

Q: What did he found?
A: the Berlin Society of Anthropology
R: the Berlin Society of Anthropology (1869) founded by Rudolph Virchow

Q: In what year?
A: 1869
R: the Berlin Society of Anthropology (1869)

Q: What was founded in 1865?
A: the Anthropological Society of Madrid
R: the Anthropological Society of Madrid (1865)

Q: And in 1870?
A: the Anthropological Society of Vienna
R: the Anthropological Society of Vienna (1870)

Q: How much later was the Italian Sociaty of Anthropology and Ethnology founded?
A: One year
R: the Anthropological Society of Vienna (1870), the Italian Society of Anthropology and Ethnology (1871)

Q: Was the American Anthropological Association founded before or after that?
A: after
R: The American Anthropological Association in 1902

Q: In what year?
A: 1902
R: The American Anthropological Association in 1902

Q: Was it an evolutionist organization?
A: Yes
R: The majority of these were evolutionist
...

Figure 6: In this example, the questioner explores questions related to time.

New Jersey is a state in the Northeastern and mid-Atlantic regions of the United States. It is a peninsula, bordered on the north and east by the state of New York; on the east, southeast, and south by the Atlantic Ocean; on the west by the Delaware River and Pennsylvania; and on the southwest by the Delaware Bay and Delaware. New Jersey is the fourth-smallest state by area but the 11th-most populous and the most densely populated of the 50 U.S. states. New Jersey lies entirely within the combined statistical areas of New York City and Philadelphia and is the third-wealthiest state by median household income as of 2016.

Q: Where is New jersey located?
A: In the Northeastern and mid-Atlantic regions of the US.
R: New Jersey is a state in the Northeastern and mid-Atlantic regions of the United States

Q: What borders it to the North and East?
A: New York
R: bordered on the north and east by the state of New York;

Q: Is it an Island?
A: **No**.
R: It is a peninsula

Q: What borders to the south?
A: Atlantic Ocean
R: bordered on the north and east by the state of New York; on the east, southeast, and south by the Atlantic Ocean

Q: to the west?
A: Delaware River and Pennsylvania.
R: on the west by the Delaware River and Pennsylvania;

Q: is it a small state?
A: Yes.
R: New Jersey is the fourth-smallest state by area

Q: How many people live there?
A: **unknown**
R: N/A

Q: Do a lot of people live there for its small size?
A: Yes.
R: the most densely populated of the 50 U.S. states.

Q: Is it a poor state?
A: **No**.
R: Philadelphia and is the third-wealthiest state by median household income as of 2016.

Q: What country is the state apart of?
A: United States
R: New Jersey is a state in the Northeastern and mid-Atlantic regions of the United States
...

Figure 7: A conversation containing *No* and *unknown* as answers.