🦆 QuAC : Question Answering in Context

Eunsol Choi★♡   **He He**★◇   **Mohit Iyyer**★♣†   **Mark Yatskar**★†
Wen-tau Yih†   **Yejin Choi**♡†   **Percy Liang**◇   **Luke Zettlemoyer**♡

Allen Institute for Artificial Intelligence†   University of Washington♡
Stanford University◇   UMass Amherst♣

{eunsol,yejin,lsz}@cs.washington.edu
{hehe,pliang}@cs.stanford.edu
{mohiti,marky,scottyih}@allenai.org

Abstract

We present QuAC, a dataset for **Qu**estion **A**nswering in **C**ontext that contains 14K information-seeking QA dialogs (100K questions in total). The dialogs involve two crowd workers: (1) a student who poses a sequence of freeform questions to learn as much as possible about a hidden Wikipedia text, and (2) a teacher who answers the questions by providing short excerpts from the text. QuAC introduces challenges not found in existing machine comprehension datasets: its questions are often more open-ended, unanswerable, or only meaningful within the dialog context, as we show in a detailed qualitative evaluation. We also report results for a number of reference models, including a recently state-of-the-art reading comprehension architecture extended to model dialog context. Our best model underperforms humans by 20 F1, suggesting that there is significant room for future work on this data. Dataset, baseline, and leaderboard available at http://quac.ai.

1 Introduction

In information-seeking dialog, students repeatedly ask teachers questions to learn about a topic of interest (Stede and Schlangen, 2004). Modeling such conversations is challenging, as the questions can be highly context-dependent, elliptical, and even unanswerable. To enable learning from rich information-seeking dialog, we present QuAC (henceforth 🦆), a large-scale dataset for **Qu**estion **A**nswering in **C**ontext that contains 14K crowdsourced QA dialogs (100K total QA pairs).[1]

Figure 1 shows an example 🦆 dialog. The interaction is student driven and centered around a short evidence text (a section from Daffy Duck's

---

[1] We use "dialog" to refer to a sequence of QA pairs.
★ Authors contributed equally.

Section: 🦆 Daffy Duck, Origin & History

STUDENT: **What is the origin of Daffy Duck?**
TEACHER: ↪ first appeared in Porky's Duck Hunt
STUDENT: **What was he like in that episode?**
TEACHER: ↪ assertive, unrestrained, combative
STUDENT: **Was he the star?**
TEACHER: ⇥ No, barely more than an unnamed bit player in this short
STUDENT: **Who was the star?**
TEACHER: ↛ No answer
STUDENT: **Did he change a lot from that first episode in future episodes?**
TEACHER: ↪ Yes, the only aspects of the character that have remained consistent (...) are his voice characterization by Mel Blanc
STUDENT: **How has he changed?**
TEACHER: ↪ Daffy was less anthropomorphic
STUDENT: **In what other ways did he change?**
TEACHER: ↪ Daffy's slobbery, exaggerated lisp (...) is barely noticeable in the early cartoons.
STUDENT: **Why did they add the lisp?**
TEACHER: ↪ One often-repeated "official" story is that it was modeled after producer Leon Schlesinger's tendency to lisp.
STUDENT: **Is there an "unofficial" story?**
TEACHER: ↪ Yes, Mel Blanc (...) contradicts that conventional belief
. . .

Figure 1: An example dialog about a Wikipedia section. The student, who does not see the section text, asks questions. The teacher provides a response in the form of a text span (or No answer), optionally yes or no (Yes / No), and encouragement about continuing a line of questioning (should, ↪ , could ⇥ , or should not ↛ ask a follow-up question).

Wikipedia page), which only the teacher can access. Given just the section's heading, "Origin & History", the student aims to learn as much as possible about its contents by asking questions. The teacher answers these questions with spans from the evidence text, as in existing reading comprehension tasks (Rajpurkar et al., 2016). Additionally, the teacher uses dialog acts to provide the student with feedback (e.g., "ask a follow up ques-

Dataset	Multi turn	Text-based	Dialog Acts	Simple Evaluation	Unanswerable Questions	Asker Can't See Evidence
🦆 **QuAC**	✔	✔	✔	✔	✔	✔
CoQA (Reddy et al., 2018)	✔	✔	✘	✔	✔	✘
CSQA (Saha et al., 2018)	✔	✘	✘	✘	✔	✘
CQA (Talmor and Berant, 2018)	✔	✔	✘	✔	✘	✔
SQA (Iyyer et al., 2017)	✔	✘	✘	✔	✘	✘
NarrativeQA (Kociský et al., 2017)	✘	✔	✘	✘	✘	✔
TriviaQA (Joshi et al., 2017)	✘	✔	✘	✔	✘	✔
SQuAD 2.0 (Rajpurkar et al., 2018)	✘	✔	✘	✔	✔	✘
MS Marco (Nguyen et al., 2016)	✘	✔	✘	✘	✔	✔
NewsQA (Trischler et al., 2016)	✘	✔	✘	✔	✔	✔

Table 1: Comparison of the QUAC dataset to other question answering datasets.

tion"), which makes the dialogs more productive.

We collect the dataset in an interactive setting where two crowd workers play the roles of teacher and student. To encourage natural and diverse questions, we do not follow previous dialog-style QA datasets that semi-automatically generate questions (Talmor and Berant, 2018; Saha et al., 2018). Furthermore, unlike QA datasets such as SQuAD and CoQA (Reddy et al., 2018), students in 🦆 do not know the answers to their questions prior to asking them, which lessens the role of string matching and simple paraphrasing in answering their questions. This property makes 🦆 similar to datasets that contain real user queries on search engines (Nguyen et al., 2016).

🦆 contains many challenging phenomena unique to dialog, such as coreference to previous questions and answers and open-ended questions that must be answered without repeating previous information (Section 3). Additionally, despite lacking access to the section text, we find that students start dialogs by asking questions about the beginning of the section before progressing to asking questions about the end. These observations imply that models built for 🦆 must incorporate the dialog context to achieve good performance.

We present a strong neural baseline (Clark and Gardner, 2018) that considers both dialog context and section text. While this model achieves within 6 F1 of human performance on SQuAD, it performs 20 F1 points below the human upper bound on 🦆, indicating room for future improvement.

2 Dataset collection

This section describes our data collection process, which involves facilitating QA dialogs between crowd workers. Table 1 shows 🦆 shares many of the same positive characteristics of existing QA datasets while expanding upon the dialog aspect.

	Train	Dev.	Test	Overall
questions	83,568	7,354	7,353	98,407
dialogs	11,567	1,000	1,002	13,594
unique sections	6,843	1,000	1,002	8,854
tokens / section	396.8	440.0	445.8	401.0
tokens / question	6.5	6.5	6.5	6.5
tokens / answer	15.1	12.3	12.3	14.6
questions / dialog	7.2	7.4	7.3	7.2
% yes/no	26.4	22.1	23.4	25.8
% unanswerable	20.2	20.2	20.1	20.2

Table 2: Statistics summarizing the 🦆 dataset.

2.1 Interactive Task

Our task pairs up two workers, a *teacher* and a *student*, who discuss a section s (e.g., "Origin & History" in the example from Figure 1) from a Wikipedia article about an entity e (Daffy Duck). The student is permitted to see only the section's title t and the first paragraph of the main article b, while the teacher is additionally provided with full access to the section text.

The task begins with the student formulating a free-text question q from the limited information they have been given. The teacher is not allowed to answer with free text; instead, they must select a contiguous span of text defined by indices (i, j) into the section text s.[2] While this decision limits the expressivity of answers, it makes evaluation simpler and more reliable; as such, it has been adopted in other reading comprehension datasets such as SQuAD, TriviaQA (Joshi et al., 2017), and NewsQA (Trischler et al., 2016).

To facilitate more natural interactions, teachers must also provide the student with a list of dialog acts v that indicates the presence of any of n discrete statements. We include three types of di-

---

[2]We set the maximum answer length to 30 tokens to prevent teachers from revealing the full article all at once.

Figure 2: A treemap visualization of the eight most frequent "Wh" words in 🦆, where box area is proportional to number of occurrences. Compared to other machine comprehension datasets, we observe increased contextuality and open-endedness, as well as a variety of both general and specific questions.

alog acts: (1) continuation (`follow up`, `maybe follow up`, or `don't follow up`), (2) affirmation (`yes`, `no`, or `neither`) and (3) answerability (`answerable` or `no answer`). The continuation act is crucial for workers to have productive dialogs, as it allows teachers to guide the student's questioning towards aspects of the article that are especially important or interesting. Altogether, a teacher's complete answer to a question q includes a pair of indices and dialog indicators, $a = (i, j, v)$. If a question is marked `no answer`, the indices are ignored.

After receiving an answer from the teacher, the student asks another question. At every turn, the student has more information about the topic than they did previously, which encourages them to ask follow-up questions about what they have just learned. The dialog continues until (1) twelve questions are answered, (2) one of the partners decides to end the interaction, or (3) more than two unanswerable questions were asked.

2.2 Collection Details

We used Amazon Mechanical Turk for collection, restricting the task to workers in English-speaking countries and with more than 1000 HITs with at least a 95% acceptance rate. We paid workers per the number of completed turns in the dialog, which encourages workers to have long dialogs with their partners, and discarded dialogs with less than three

QA pairs.[3] To ensure quality, we created a qualification task and allowed workers to report their partner for various problems. More details on data collection can be found in our datasheet.[4]

Article selection Our early pilot studies showed that articles about people generally require less background knowledge to write good questions than other categories. To find articles about people with varied backgrounds, we retrieved articles from a list of category keywords (culture, animal, people associated with event, geography, health, celebrity) using a web interface provided by the Wikimedia foundation.[5] We pruned by popularity by selecting articles with at least 100 incoming links, and we additionally removed non-person entities using YAGO (Suchanek et al., 2007). After article selection, we filtered sections from these articles based on the number of paragraphs, number of tokens, and average words per sentence. [6]

Dataset validation To create our evaluation sets, we collected four additional annotations per question. Workers were presented with questions from a previously collected dialog and asked to

provide answer spans.[7] Acquiring many annotations is important since many questions in 🛶 have multiple valid answers.

Train / Dev / Test Differences Table 2 shows small differences between training, development and testing splits. Sections in the training set are shorter than those in the evaluation folds because we permit multiple dialogs about the same section only in training; since workers preferred reading shorter sections, these were more likely to result in multiple dialogs. Variations in answer span length arise from two sources: (1) having multiple annotations in the validation task and (2) differing incentives between the data collection and validation procedures.[8] An analysis measuring the effect of these variations shows that they result in little difference in evaluation.[9]

3 Dataset Analysis

🛶 differs from other reading comprehension datasets due to our dialog-style collection process and the information asymmetry between teacher and student. In the following sections, we provide a qualitative analysis of the dataset in 🛶 that highlights challenging question types as well as the impact of the dialog context.

Question and answer types Table 2 shows dataset summary statistics. 🛶 has long answers of 15 tokens on average compared to 3 for SQuAD, which is unsurprising as most SQuAD answers are either entities or numerics (Jurczyk et al., 2018) while 🛶 questions can be more open-ended. While the average question length (6.5 tokens) is shorter than that of SQuAD (11 tokens), this does not indicate reduced question complexity, as the student (1) cannot access the section to paraphrase it and (2) can be more concise by coreferencing previous interactions.

Figure 2 visualizes the most frequent question types in 🛶 based on "Wh" words.[10] For a more

Figure 3: An example successful dialog from 🛶. Questions build on each other and interesting aspects (e.g., plagiarism) are explored as they are discovered.

fine-grained analysis, we randomly sampled 100 questions (each from a different dialog) and manually labeled different phenomena in Table 3. Unlike most current QA datasets that focus on factoid questions, our task setup encourages more open-ended questions: about half of 🛶 questions are non-factoid. Furthermore, 86% of questions are contextual, requiring reading the context to resolve coreference; of these, 44% refer to entities or events in the dialog history, while 61% refer to the subject of the article.

The role of context Dialog context is crucial to understanding and answering 🛶 questions. Figure 5a shows that the location of the answer within the text is influenced by the number of questions asked previously. Early questions are mostly answered in the beginning of the section, while later questions tend to focus on the end of the section. Interestingly, text in the middle of the section is not asked about as frequently (Figure 5c). As more questions get asked, the more likely a question is to be unanswerable.

Figure 5b shows how the answers progress through different chunks of the evidence text (where each section is divided into 12 chunks of

---

[7] After submitting an answer, they were shown the original teacher's answer so that they could understand the context of the subsequent questions.

[8] Validation workers did not have to maintain the dialog and so did not include as much information in the response.

[9] More specifically, we analyze whether references from the initial data collection significantly differ from references collected during validation. We observe a difference of less than 1 F1 when using the original answer as system output versus using validation answers.

[10] To more effectively visualize sub-boxes like "what did", we exclude questions from the tail of the distribution.

Section: Gaelic Ireland : Invasion

STUDENT: **What year did the invasion happen?**

TEACHER: ↪ in 1169 the main body of Norman, Welsh and Flemish forces landed in Ireland and quickly retook Leinster and the cities of Waterford and Dublin on behalf of Diarmait.

STUDENT: **Who was Diarmait?**

TEACHER: ⇨ King Diarmait Mac Murchada of Leinster.

STUDENT: **Where is Leinster located?**

TEACHER: ↛ landed in Ireland and quickly retook Leinster.

STUDENT: **Were invasions common?**

TEACHER: ↛ No answer

STUDENT: **Are there any other interesting aspects about this article?**

TEACHER: ⇨ Yes, IPope Adrian IV, the only English pope, had already issued a Papal Bull in 1155 giving Henry II of England authority to invade Ireland.

STUDENT: **Who lead the invasion?**

TEACHER: ↛ No answer

STUDENT: **Did England defeat the Irish armies?**

TEACHER: ↛ No answer

Figure 4: A less successful dialog from 🥞. The student struggles to get information despite asking good questions. The teacher attempts to provide extra context to guide the student, but the dialog ultimately ends because of too many unanswerable questions.

equal size). The answer to the next question is most frequently either in the same chunk as the previous question or an adjacent chunk, and most dialogs in the dataset cover three to six of the chunks (Figure 5d). These observations suggest that models for 🥞 must take into account the dialog context. However, results in Section 5 show that solely relying on the location of previous answers is not sufficient.

Finally, we examine properties of the questions as a function of the turn position in the dialog (Figure 6). The frequency of yes/no questions increases significantly as the dialogs progress; again, at the beginning of the dialog, students have very little information, so it is harder to formulate a yes/no question. The percentage of questions that have multiple answers declines as the dialog progresses, implying students ask general questions first and specific ones later.

Qualitative examples Figures 3 and 4 contain two representative dialogs from 🥞. Longer dialogs sometimes switch topics (such as in Figure 3 about "academic work") and often go from general to specific questions. Students whose ques-

Question type	%	Example
Non-factoid	54	Q: **Were** the peace talks **a success?** Q: **What was** her childhood **like?**
Contextual	86	
Coref (article)	61	Title: Paul Cézanne: Early years Q: When did **he** start painting?
Coref (history)	44	Q: What was special about the Harrah's? A: project was built by Trump with financing from the Holiday Corporation. Q: **Which led to what?**
Anything else?	11	Q: What **other acting** did he do? Q: **What else** did he research?

Table 3: An analysis of 🥞 questions. **Non-factoid** questions do not ask about specific facts, while **contextual** questions require reading the history to resolve coreferences to the dialog history and/or article.

tions go unanswered commonly resort to asking their teacher for any interesting content; even if this strategy fails to prolong the dialog as in Figure 4, models can still use the dialog to learn when to give no answer.

4 Experimental Setup

We consider the following QA task: given the first k questions and k ground-truth answers in the dialog, all supporting material (entity e, topic t, background b, and section text s), and question q_{k+1}, we predict the answer span indices i, j in the section text s. Since affirmation questions are incomplete without a yes/no answer and the continuation feedback is important for information-seeking dialog, we predict the dialog acts v, which with the span form the final answer prediction a_{k+1}.

All of our experiments are carried out on a train/dev/test split of 83.5k/7.3k/7.3k questions/answer pairs, where no sections are shared between the different folds. Questions in the training set have one reference answer, while dev and test questions have five references each. For all experiments, we do not evaluate on questions with a human F1 lower than 40, which eliminates roughly 10% of our noisiest annotations.[11]

4.1 Evaluation Metrics

Our core evaluation metric, word-level F1, is implemented similarly to SQuAD (Rajpurkar et al.,

___
[11]A manual inspection of annotations below this threshold revealed many lower quality questions; however, we also report unthresholded F1 in the final column of Table 4.
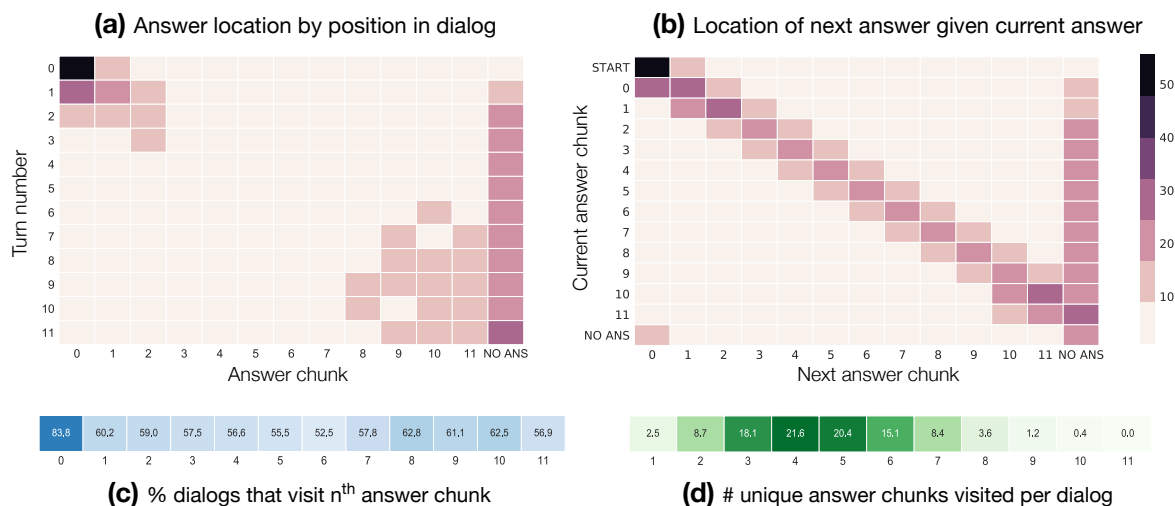
(a) Answer location by position in dialog

(b) Location of next answer given current answer

(c) % dialogs that visit n^{th} answer chunk

(d) # unique answer chunks visited per dialog

Figure 5: Heatmaps depicting the importance of context in 🦜 dialogs, where (a) and (b) share the same color scale. The student's earlier questions are answered mostly by the first few chunks, while the end of the section is covered in later turns (a). The middle is the least covered portion (c), and dialogs cover around five unique chunks of the section on average (d). The transition matrix (b) shows that the answer to the next question is more likely to be located within a chunk adjacent to the current answer than in one farther away.
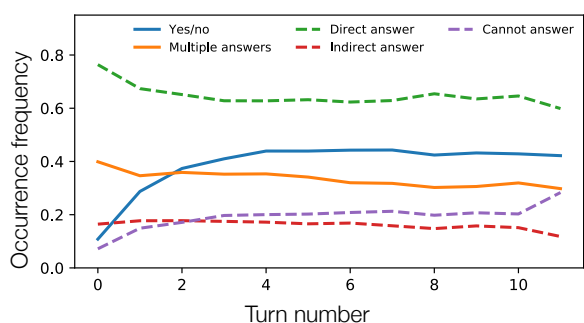

Figure 6: The number of turns in the dialog influences the student's behavior: they start by asking general questions (i.e., easier to answer, with multiple possible answers) and progress to more specific ones.

2016): precision and recall are computed by considering the portion of words in the prediction and references that overlap after removing stopwords.[12] For `no answer` questions, we give the system an F1 of one if it correctly predicts `no answer` and zero otherwise.[13] Like SQuAD, we compute the maximum F1 among all references; however, since many 🦜 questions have multiple valid answers, this metric varies significantly with

the number of reference annotations. To make oracle human and system performance comparable, given n references, we report the average of the maximum F1 computed from each $n-1$ subset with respect to the heldout reference.

Additionally, since averaged F1 can be misleading for questions with multiple valid answers, we introduce the human equivalence score (HEQ), a performance measure for judging whether a system's output is as good as that of an average human.[14] HEQ measures the percentage of examples for which system F1 exceeds or matches human F1. We compute two variants: (1) the percentage of questions for which this is true (HEQ-Q), and (2) the percentage of dialogs for which this is true for every question in the dialog (HEQ-D). A system that achieves a value of 100 on HEQ-D can by definition maintain average human quality output over full dialogs.

For dialog acts, we report accuracy with respect to the majority annotation, breaking ties randomly.

5 Experiments

5.1 Sanity checks

Random sentence This baseline selects a random sentence in the section text s as the answer (including `no answer`).

---

[12]Since our answer spans have vaguer boundaries than the shorter ones in SQuAD, exact match is not a useful metric.

[13]Because the validation task was more susceptible to spam by constant annotation of "no-answer," we only allow "no-answer" if the majority of references marked "no-answer", removing other answers. If "no-answer" is not the majority answer, we remove all instances of "no-answer".

[14]In cases with lower human agreement on F1, if a system produces one reference exactly (F1 = 100), it will get points that it can use to offset poor performance on other examples.

Majority The majority answer outputs `no answer` and the majority class for all other dialog acts (`neither` for affirmation and `don't follow up` for continuation).

Transition matrix We divide the supporting text into 12 chunks (with a special chunk for `no answer`) and use the transition matrix (computed from the training set) in Figure 5b to select an answer given the position of the previous answer. This baseline does not output other dialog acts.

5.2 Upper bounds

Gold NA + TM This is the same transition matrix (TM) baseline as before, except that for questions whose gold annotations are `no answer`, we always output `no answer`.

Gold sentence + NA To see if 🥏 can be treated as an answer sentence selection problem, we output the sentence from s with the maximal F1 with respect to references, or `no answer` for unanswerable questions.

Human performance We pick one reference as a system output and compute the F1 with respect to the remaining references using the method described in Section 4.1. By definition, all HEQ measures are 100, and we report agreement for the affirmation dialog act.[15]

5.3 Baselines

Pretrained InferSent To test the importance of lexical matching in our dataset, we output the sentence in s whose pretrained InferSent representation (Conneau et al., 2017) has the highest cosine similarity to that of the question.

Feature-rich logistic regression We train a logistic regression using Vowpal Wabbit (Langford et al., 2007) to select answer sentences. We use simple matching features (e.g., n-gram overlap between questions and candidate answers), bias features (position and length of a candidate), and contextual features (e.g., matching features computed with previous questions / answers, turn number).

BiDAF++ We use a re-implementation of a top-performing SQuAD model (Peters et al., 2018) that augments bidirectional attention flow (Seo

et al., 2016, BiDAF) with self-attention (Clark and Gardner, 2018) and contextualized embeddings.[16]

A token for `no answer` is appended to s to enable its prediction following Levy et al. (2017). Additionally, we modify the model for our task to also predict dialog acts, placing a classifier over the same representation used to predict the end position of the predicted span.

BiDAF++ w/ k-ctx As BiDAF++ does not model any dialog context, we modify the passage and question embedding processes to consider the dialog history. We consider context from the previous k QA pairs.[17]

- **Passage embedding** We explicitly identify the previous k answers within the section text by concatenating marker embeddings to the existing word embeddings.

- **Question embedding** Naively prepending the previous k questions to the current question did not show gains in initial experiments. We opt instead to simply encode the dialog turn number within the question embedding.

5.4 Results

Table 4 summarizes our results (each cell displays dev/test scores), where dialog acts are Yes/No (affirmation) and Follow up (continuation). For comparison to other datasets, we report F1 without filtering low-agreement QA pairs (F1').

Sanity check Overall, the poor sanity check results imply that 🥏 is very challenging. Of these, following the transition matrix (TM) gives the best performance, reinforcing the observation that the dialog context plays a significant role in the task.

Upper bounds The human upper bound (80.8 F1) demonstrates high agreement. While Gold sentence + NA does perform well, indicating that significant progress can be made by treating the problem as answer sentence selection, HEQ measures show that span-based approaches will be needed achieve average human equivalence. Finally, the Gold NA + TM shows that 🥏 cannot be solved by ignoring question and answer text.

---

[15]We did not collect multiple annotations for the continuation dialog act and so omit it.

[16]The AllenNLP (Gardner et al., 2017) implementation we use reaches 82.7 on the SQuAD development set, compared to the paper's reported 85.8 on SQuAD; regardless, this implementation would have been state-of-the-art less than a year ago, making it an extremely strong baseline.

[17]Our implementation is available in AllenNLP.

	F1	HEQ-Q	HEQ-D	Yes / No	Follow up	F1 (All)
Random sentence	15.7 / 15.6	6.9 / 6.9	0.0 / 0.1	—	—	16.4 / 16.3
Majority answer	22.7 / 22.5	22.7 / 22.5	0.5 / 0.4	78.8 / 77.6	57.9 / 56.7	20.2 / 20.0
Trans. matrix (TM)	31.8 / 31.5	15.8 / 15.8	0.1 / 0.2	—	—	31.2 / 30.9
Pretrained InferSent	21.4 / 20.8	10.2 / 10.0	0.0 / 0.0	—	—	22.0 / 21.4
Logistic regression	34.3 / 33.9	22.4 / 22.2	0.6 / 0.2	—	—	34.3 / 33.8
BiDAF++ (no ctx)	51.8 / 50.2	45.3 / 43.3	2.0 / 2.2	86.4 / 85.4	59.7 / 59.0	50.1 / 48.2
BiDAF++ (w/ 1-ctx)	59.9 / 59.0	54.9 / 53.6	4.7 / 3.4	86.5 / **86.1**	61.3 / 60.3	57.5 / 56.5
BiDAF++ (w/ 2-ctx)	**60.6 / 60.1**	**55.7 / 54.8**	**5.3** / 4.0	**86.6** / 85.7	**61.6 / 61.3**	**58.3 / 57.8**
BiDAF++ (w/ 3-ctx)	**60.6** / 59.5	55.6 / 54.5	5.0 / **4.1**	86.1 / 85.7	**61.6** / 61.2	58.1 / 57.0
Gold NA + TM	43.0 / 42.6	27.4 / 27.4	1.0 / 0.8	—	—	41.0 / 40.6
Gold sentence + NA	72.4 / 72.7	61.8 / 62.7	9.8 / 9.7	—	—	70.8 / 71.2
Human performance	80.8 / 81.1	100 / 100	100 / 100	89.4 / 89.0	—	74.6 / 74.7

Table 4: Experimental results of sanity checks (top), baselines (middle) and upper bounds (bottom) on 🦆. Simple text matching baselines perform poorly, while models that incorporate the dialog context significantly outperform those that do not. Humans outperform our best model by a large margin, indicating room for future improvement.

Baselines Text similarity methods such as bag-of-ngrams overlap and InferSent are largely ineffective on 🦆, which shows that questions have little direct overlap with their answers. On the other hand, BiDAF++ models make significant progress, demonstrating that existing models can already capture a significant portion of phenomena in 🦆. The addition of information from previous turns (w/ 1-ctx) helps significantly, indicating that integration of context is essential to solving the task. While increasing the context size in BiDAF++ continues to help, we observe saturation using contexts of length 3, suggesting that more sophisticated models are necessary to take full advantage of the context. Finally, even our best model underperforms humans: the system achieves human equivalence on only 60% of questions and 5% of full dialogs.

5.5 Error Analysis

In this section, we analyze the development set performance of our best context-aware model (BiDAF++ w/ 2-ctx), our best context-agnostic model (BiDAF++), and humans. Figure 7 contains three plots showing how F1 scores of baseline models and human agreement vary with (1) turn number, (2) distance from previous answer,[18] and (3) answer length in tokens. Taken as a whole, our analysis reveals significant qualitative differences between our context-aware and context-agnostic models beyond simply F1; additionally, human

behavior differs from that of both models.

In the first plot, human agreement is unchanged throughout the dialog while the performance of both models decreases as the number of turns increases, although the context-aware model degrades less. While continuing a dialog for more turns does not affect human agreement, the second plot shows that human disagreement increases as the distance between the current answer's location within the section text and that of the previous answer increases. Larger distances indicate shifts in the student's line of questioning (e.g., if the teacher told the student not to follow up on the previous question). The plot also shows that model performance suffers (significantly more than humans) as distance increases, although the context-aware model can tolerate smaller shifts better than the context-agnostic model. In the last plot, human agreement is higher when the answer span is short; in contrast, our model struggles to pin down short answers compared to longer ones.

The plots demonstrate the increased robustness of the context-aware model compared to BiDAF++. This finding is reinforced by examining the difference in model performance on questions where previously the teacher recommended the student to "follow up" vs. not to follow up. The context-aware baseline performs 6 HEQ-Q higher on the "follow up" questions; in contrast, the context-agnostic baseline shows no HEQ-Q difference between the two types of questions. This discrepancy stems from the context-agnostic
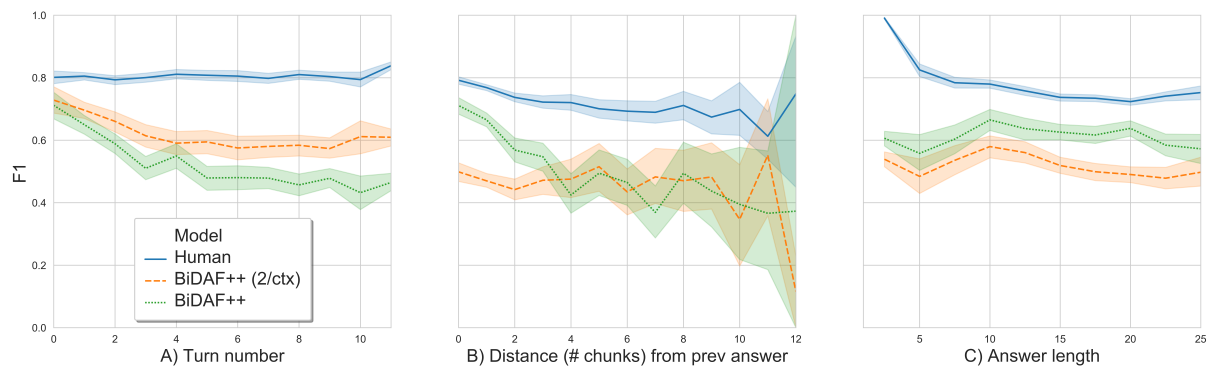
---

[18]We divide the text into 12 equally-sized chunks and compute the difference of the current and previous chunk indices.

Figure 7: The F1 scores of baseline models and human agreements based on dialog turn number, answer's distance from previous answer, and the answer span token length.

model's inability to take advantage of the location of the previous answer.

6 Related Work

Reading Comprehension Our work builds on span based reading comprehension (Rajpurkar et al., 2016; Joshi et al., 2017; Trischler et al., 2016), while also incorporating innovations such as curating questions independently of supporting text to reduce trivial lexical overlap (Joshi et al., 2017; Kociský et al., 2017) and allowing for unanswerable questions (Trischler et al., 2016; Rajpurkar et al., 2018). We handle open-ended questions like in MSMARCO (Nguyen et al., 2016), with multiple references, but we are the first to incorporate these into information-seeking dialog.

Sequential QA Our work is similar to sequential question answering against knowledge bases (Iyyer et al., 2017) and the web (Talmor and Berant, 2018), but instead of decomposing a single question into smaller questions, we rely on the curiosity of the student to generate a sequence of questions. Such open information seeking was studied in semantic parsing on knowledge bases (Dahl et al., 1994) and more recently with modern approaches (Saha et al., 2018), but with questions paraphrased from templates. Concurrent to our work, Saeidi et al. (2018) proposed a task of generating and answering yes/no questions for rule focused text (such as traffic laws) by interacting with a user through dialog. Also concurrently, Reddy et al. (2018) propose conversational question answering (CoQA) from text but allow both students and questioners to see the evidence. As a result, a large percentage of CoQA answers are named entities or short noun phrases, much like those in SQuAD. In contrast, the asymmetric nature of 🦆 forces students to ask more

exploratory questions whose answers can be potentially be followed up on.[19]

Dialog 🦆 fits into an increasing interest in open domain dialog, mostly studied in the context of social chit-chat (Li et al., 2016; Ritter et al., 2011; Fang et al., 2017; Ghazvininejad et al., 2018). Most related to our effort is visual dialog (Das et al., 2017), which relies on images as evidence instead of text. More explicit goal driven scenarios, such as bargaining (Lewis et al., 2017) and item guessing (He et al., 2017) have also been explored, but the language is more constrained than in 🦆. Information-seeking dialog specifically was studied in Stede and Schlangen (2004).

7 Conclusion

In this paper, we introduce 🦆, a large scale dataset of information-seeking dialogs over sections from Wikipedia articles. Our data collection process, which takes the form of a teacher-student interaction between two crowd workers, encourages questions that are highly contextual, open-ended, and even unanswerable from the text. Our baselines, which include top performers on existing machine comprehension datasets, significantly underperform humans on 🦆. We hope this discrepancy will spur the development of machines that can more effectively participate in information seeking dialog.

Acknowledgments

---

[19]On average, CoQA answers are 2.7 tokens long, while SQuAD's are 3.2 tokens and 🦆 's are over 14 tokens.

References

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. *Proceedings of the Association for Computational Linguistics*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of Empirical Methods in Natural Language Processing*.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William M. Fisher, Kate Hunicke-Smith, David S. Pallett, Christine Pao, Alexander I. Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *Proceedings of the workshop on Human Language Technology*.

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*.

Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding board–university of washingtons alexa prize submission. *Alexa Prize Proceedings*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. Allennlp: A deep semantic natural language processing platform.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. *Association for the Advancement of Artificial Intelligence*.

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. *Proceedings of the Association for Computational Linguistics*.

Mohit Iyyer, Wen tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the Association for Computational Linguistics*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Association for Computational Linguistics*.

Tomasz Jurczyk, Amit Deshmane, and Jinho D. Choi. 2018. Analysis of wikipedia-based corpora for question answering. *arXiv preprint arXiv:abs/1801.02073*.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, abs/1712.07040.

John Langford, Lihong Li, and Alex Strehl. 2007. Vowpal wabbit online learning project.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke S. Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Conference on Computational Natural Language Learning*.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *Proceedings of Empirical Methods in Natural Language Processing*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *Proceedings of Empirical Methods in Natural Language Processing*.

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*, abs/1611.09268.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Conference of the North American Chapter of the Association for Computational Linguistics*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the Association for Computational Linguistics*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of Empirical Methods in Natural Language Processing*.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *ArXiv*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of Empirical Methods in Natural Language Processing*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Signh, Tim Rocktschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of Empirical Methods in Natural Language Processing*.

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*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *Proceedings of the International Conference on Learning Representations*.

Manfred Stede and David Schlangen. 2004. Information-seeking chat: Dialogues driven by topic-structure. In *Eighth workshop on the semantics and pragmatics of dialogue; SemDial*.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the World Wide Web Conference*.

A. Talmor and J. Berant. 2018. The web as knowledge-base for answering complex questions. In *Conference of the North American Chapter of the Association for Computational Linguistics*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.