Datasheet for (QuAC)

1 Motivation for Datasheet Creation

Why was the dataset created?
We collected to facilitate designing and evaluating models for information-seeking dialog, a sequential QA task that involves resolving coreferences, dealing with unanswerable questions, and leveraging world knowledge.

Has the dataset been used already?
All papers reporting on are required to submit their results to http://quac.ai.

Who funded the dataset?
was co-funded by the Allen Institute of Artificial Intelligence and the DARPA CwC program through ARO (W911NF-15-1-0543).

2 Dataset Composition

What are the instances?
The core problem involves predicting a text span to answer a question about a Wikipedia section. Since questions include a dialog component, each instance includes a "dialog history" of questions and answers asked in the dialog prior to the given question, along with some additional metadata.

How many instances are there?
contains 98,407 QA pairs from 13,594 dialogs. The dialogs were conducted on 8,854 unique sections from 3,611 unique Wikipedia articles, and every dialog contains between four and twelve questions.

What data does each instance consist of?
The instances come from a QA dialog about a given section of a Wikipedia article. Each instance is a tuple of ⟨question q_i from a dialog, preceding questions $q_{1...i-1}$ and their associated answers $a_{1...i-1}$, first paragraph of Wikipedia article, text of Wikipedia section, article title, section title ⟩ and the output is the answer a_i to question q_i. In addition to spans of text from the section, answers also include dialog acts: *affirmation* for yes/no questions; *continuation*, which specifies whether the current line of questioning should be continued or not; and an *unanswerable flag* for questions that cannot be answered from the section text.

Does the data rely on external resources?
No, everything is included in our release.

Are there recommended data splits or evaluation measures?
The release comes with a train/dev split such that there is no overlap in sections across splits. Furthermore, the dev and test sets only include one dialog per section, in contrast to the training set which can have multiple dialogs per section. Dev and test instances come with five reference answers instead of just one as in the training set; we obtain the extra references to improve the reliability of our evaluations, as questions can have multiple valid answer spans. The test set is not publicly available; instead, researchers must submit their models to the leaderboard at http://quac.ai, which will run the model on our hidden test set.

We provide an official evaluation script for used by our leaderboard for test set evaluation. The script computes two metrics: word-level F1 and human equivalence (HEQ). If a particular instance has n reference annotations, we compute F1 by averaging the maximum F1 over all subsets of $n-1$ annotators. We compute two flavors of human equivalence: HEQ-Q, which simply measures the percentage of questions for which the system's F1 matches or exceeds human F1, and HEQ-D, which measures the percentage of *dialogs* where all questions within have an HEQ-Q $= 1$.

3 Data Collection Process

How was the data collected?
The Wikipedia articles were filtered to those in the "people" category associated with various subcategories (culture, animal, people associated with event, geography, health, celebrity) and downloaded using a web interface provided by the Wikimedia Foundation (http://petscan.wmflabs.org). The dialogs were collected using Amazon Mechanical Turk.

Who was involved in the collection process and what were their roles?
We (the authors) did initial pilot studies amongst ourselves to refine the data collection and validation tasks. After multiple rounds of pilots, we launched the large scale data collection via Amazon Mechanical Turk.

In this task, each dialog is an interaction between two crowd workers (a *teacher* and a *student*) paired up in a chat room. The teacher has ac-

cess to the full text of a section from a Wikipedia article and uses it to answer questions from the student, who has only minimal information about the section (i.e., its title and a short summary of the article). The teacher cannot write free text answers; instead, they must answer with spans of text from the section. To reduce lag time (e.g., the teacher doing nothing while waiting for the student's next question), we have each worker perform both roles simultaneously on two different sections. Finally, we terminate dialogs either when they reach a maximum of 12 QA pairs or when 3 unanswerable questions have been asked (as many unanswerable questions often signal an unproductive dialog).

We set up a qualification task to filter out spammers and workers who provided low-quality responses. 278 workers passed the task and ended up contributing to , compared to 367 workers who were rejected. Workers were paid per question on a variable pay scale (e.g., the eighth question of a dialog paid more than the first question); on average, we paid each worker $6 per hour during the dialog collection phase.

Finally, as previously mentioned, we gather multiple annotations for our dev and test sets with a validation task. This task is designed differently from the main collection task: a single worker acts as the teacher, answering pre-recorded student questions from the main collection task. After answering each question, the validation worker is provided with the original teacher's answer in order to make sense of subsequent questions in the dialog. The difference in task design contributes to discrepancies between dev and test answer lengths compared to train answer lengths (Table 1). Validation workers are paid a fixed rate for each question they answer, while teachers in the main task earn more for long dialogs than short ones. Thus, teachers are incentivized to provide longer answer spans than validation workers, as the additional information they provide might trigger the student to ask questions that they would not have asked otherwise. Articles are shorter in train because it is the only fold where we allow multiple dialogs about the same section, and workers preferred to work on shorter sections than longer ones.[1]

Over what time frame was the data collected?

The data was collected over a two week period.

[1]Sections with multiple dialogs are almost 60 tokens shorter on average than sections with just a single dialog.

	Train	Dev.	Test	Overall
unique sections	6,843	1,000	1,002	8,854
dialogs	11,567	1,000	1,002	13,594
questions	83,568	7,354	7,353	98,407
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 1: Statistics summarizing the dataset.

Does the dataset contain all possible instances?

No. Information-seeking dialog can be conducted over any topic, not just about people. We restricted to people entities in certain subcategories because our pilot experiments showed that crowd workers had an easier time asking questions about people than about arbitrary topics, in large part because many topics require specialized knowledge to intelligently converse about.

Furthermore, we subsampled the "people" articles to only popular articles (measured by number of incoming links), which biases in favor of celebrities (e.g., musicians, politicans). We added this filtering step after observing that many articles were poorly written and did not contain enough details to result in long and cohesive dialogs. Since more popular articles tend to have more content moderation, they are better suited for our task. We also had length filters on the articles: only sections between 250 and 550 words that contained between 3 and 5 paragraphs were selected for the crowdsourced task. These thresholds were refined through pilot studies to balance worker fatigue (longer sections take more time and effort to read) with task feasibility (it is very difficult to ask good questions when provided with too little information).

Finally, the way in which we collected our data is a simplification of real-world information-seeking dialog. We disallowed teachers from providing free-text responses, which limits the type of feedback they can provide the student. We made this decision because evaluating span-based answers is much more well-defined than evaluating free text; metrics such as F1 would not be usable had we allowed free text. To reclaim some of the teacher's flexibility, we added in some simple dialog acts.

If the dataset is a sample, then what is the population?
The articles used in  are not representative of the overall population of Wikipedia articles about people entities. After restricting articles based on the various criteria detailed above, we are left with a biased sample. We perform a preliminary analysis of gender bias (by counting pronouns in each article) and occupational bias in  (by counting words that are strongly associated with particular occupations; e.g., {guitarist, opera, albums} $\rightarrow$ musician).

While this analysis is obviously imperfect, its results give a rough estimate of the data composition. For gender, 76% of the articles are about men, 16% are about women, and the remaining 8% are mostly about multiple people (e.g., bands). Regarding occupations, there is a substantial bias towards entertainers: 36% of  articles are about musicians, 15% about sports figures, and 13% about movie or television personalities. The remaining 36% encompasses all other occupations.

A full analysis of all Wikipedia articles about people entities is beyond the scope of this document. While similar biases may exist in the full population of articles, we believe that our filtering process has significantly exaggerated them in  . Additionally, there likely exist other biases that we have not analyzed here (e.g., age, race).

4 Data Preprocessing

What preprocessing / cleaning was done?
Collected questions, answers, and Wikipedia text were tokenized using spaCy (http://spacy.io/. Additionally, we do not evaluate on any instances for which the human F1 is less than 40, which eliminates roughly 10% of our noisiest annotations.

Was the raw data saved in addition to the cleaned data?
We include all annotations (including ones that do not meet the minimum F1 threshold) in the released data splits. The evaluation script has a flag to ignore the noisy annotations.

Does this dataset collection/preprocessing procedure achieve the initial motivation?

 is indeed a dataset that allows researchers to explore phenomena specific to information-seeking dialog and, more generally, conversational question-answering systems. However, it is not suited for training general-purpose information-seeking dialog systems, as it contains only dialogs about a biased sample of people entities. Our article selection process was necessary to collect high-quality data at scale from crowd workers, but we hope that a better procedure will be developed in the future to collect dialogs about a more diverse set of topics.

5 Dataset Distribution

How is the dataset distributed?
It is available at http://quac.ai.

When was it released?
August 2018

What license (if any) is it distributed under?
 is distributed under the MIT license. Additionally, researchers that use  are requested to cite the corresponding dataset paper.

Who is supporting and maintaining the dataset?
The dataset will be maintained by the first four authors of the paper: Eunsol Choi, He He, Mohit Iyyer, and Mark Yatskar. All updates will be posted on the dataset website.

6 Legal & Ethical Considerations

Were workers told what the dataset would be used for and did they consent?
Crowd workers were not told of the specific nature of the dataset; however, they consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement.

If it relates to people, could this dataset expose people to harm or legal action?
No, the collected dialogs do not contain personal information about the crowd workers, and the Wikipedia articles were already public.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?
The way in which articles were selected for  introduced biases. All articles are about famous people, which advantages males in a small number of professions. In addition, the dataset could very well contain other forms of bias that we have not analyzed in detail. For this reason, we do not recommend that models trained on  be deployed in real-world settings.