Datasheet for Discourse Structure of Long-form Answers

Fangyuan Xu[♢] **Junyi Jessy Li**[♡] **Eunsol Choi**[♢]
[♢]Department of Computer Science
[♡]Department of Linguistics
The University of Texas at Austin
{fangyuan, jessy, eunsol}@utexas.edu

1 Motivation for Datasheet Creation

Why was the dataset created? Long-form answers, consisting of multiple sentences, can provide nuanced and comprehensive answers to a broader set of questions as compared to short-form answers. But there is little computational study on the structure of such long-form answers. We developed an ontology of six sentence-level functional roles for long-form answers – the roles we defined are **Answer**, **Answer Summary**, **Example**, **Auxiliary Information**, **Organizational Sentence** and **Miscellaneous**. We provide annotations for 3.9k sentences in 640 answer paragraphs from three pre-existing long-form question answering datasets —ELI5 (Fan et al., 2019), WebGPT (Nakano et al., 2021) and Natural Questions (NQ) (Kwiatkowski et al., 2019). Please refer to our paper (Xu et al., 2022) for detailed descriptions and examples of each role.

Has the dataset been used already? We have used the dataset to (1) analyze the discourse structure of different types of long-form answer; (2) study automatically classifying roles for sentences in long-form answer paragraphs.

Who funded the datasheet? The project is funded by NSF grants IIS-1850153, IIS-2107524 and by UT Austin.

2 Dataset Composition

What are the instances? Each instance is a (question, long-form answer) pair from one of the four data sources – ELI5, WebGPT, NQ, and model-generated answers (denoted as ELI5-model), and our discourse annotation, which consists of QA-pair level validity label and sentence-level functional role label.

The QA pairs are sourced from the validation split of ELI5 from the KILT (Petroni et al., 2021) benchmark, the testing portion of human demonstration from WebGPT [1], and the validation split from NQ. For ELI5-model, we sampled from four different model configurations reported in Krishna et al. (2021), i.e. combination of nucleus sampling threshold p={0.6, 0.9}, and generation conditioning on {predicted, random} passages.

Data	Validity	Role
ELI5	1,035 (6,575)	411 (2,670)
ELI5-model	193 (1,839)	115 (1,080)
WebGPT	100 (562)	98 (551)
NQ	263 (1,404)	131 (695)
Total	1,591 (10,380)	755 (4,996)

Table 1: Data Statistics. For validity and role, the first number in each cell corresponds to the number of long-form answers, and the second number represents the number of sentences.

What data does each instance consist of? We provide two types of data – validity data and functional role data. Please see Table 2 for example instances.

Each instance in the validity dataset consists of: a question q (string), an answer paragraph a (a sequence of sentences, each sentence is a string), a boolean value indicating whether it is valid (True/False), and a list of invalid reasons provided by annotators. The invalid reasons can be one of the following: (1) multiple question asked, (2) assumption rejected, (3) no valid answer and (4) nonsensical question and we include all invalid reasons selected by each annotators. The list will contain three empty lists if all annotators found the QA pair valid.

Each instance in the functional role dataset consists of: a question q, an answer paragraph a (a sequence of sentences, each sentence is a string), a se-

[1]The testing samples are publicly hosted at https://openaipublic.blob.core.windows.net/webgpt-answer-viewer/index.html, which answers questions from the ELI5 test set.

Field	Value
Dataset	ELI5
Question	What (Who?) Exactly Defines a Reliable News Source?
Answer	["For my own sake I try to browse news sources from all over the world.", "This tends to cut down on specific left/right narratives for a specific region and allows you to extract just the relevant facts of the story.", "Some discretion and intelligence is needed."]
Is Valid	False
Invalid reasons	[[no valid answer], [no valid answer], [no valid answer]]
Dataset	ELI5
Question	Why are dragons present in cultures all over the place?
Answer	["It appears that dragons are everywhere because the word dragon is used to describe any reptilian mythical creature.", "For example, Chinese dragons and European dragons aren't the same thing.", "European dragons are generally evil and breath fire. Chinese dragons are often benevolent and are associated with water.", "So the Chinese and Europeans haven't come up with the same creature, it's just given the same name in English."]
Roles	[Summary, Example, Example, Example, Example]
Raw role annotations	[[Summary, Summary, Summary], [Example, Example, Example], [Example, Example, Example], [Example, Example, Example], [Answer, Example, Example]]

Table 2: Example data. The first one is a validity annotation, which is marked as invalid for the reason of "no valid answer" by all three annotators. The second one is a role annotation. The "Raw role annotations" lists out individual annotations for each sentence (i.e. The last sentence is marked as "Answer" by one annotator, and as "Example" by the other two. We omit certain fields such as data id here due to space. Please refer to https://github.com/utcsnlp/lfqa_discourse for full description.

quence of final role annotations for each sentence in answer paragraph, and raw role annotations, which is a sequence of lists of raw role annotations from annotators.

How many instances are there? Table 1 contains the statistics of our annotated dataset. We collected validity annotations for 1.5K (question, answer) pairs and sentence-level role annotations for about half of them.

Does the data rely on external resources? No, all resources are included in our release.

Are there recommended data splits or evaluation measures? We release the train/validation/test split we used for our role classifier in our repository.

3 Data Collection Process

How was the data collected? Our data collection has two stages: (1) **Validity annotation** where annotators are presented a (question, answer) pair and annotate whether this pair is valid based on a set of pre-defined invalid reason. (2) **Role annotation** where annotators are presented a *valid* (question, answer) pair from the first stage and select a role for each sentence in the answer paragraph.

Who was involved in the collection process and what were their roles? We recruited crowdworkers from Amazon Mechanical Turk to perform the question validity annotation, and undergraduate students major in linguistic from our educational institution to perform the role annotation. We recruited crowdworkers who were from the USA, had a minimum approval rating of 95% and had completed at least 1000 HITs. We first qualified and then provided training materials to both groups of annotators. A total of 29 crowdworkers and 6 undergraduate students were invovled in the annotation.

Over what time frame was the data collected? The dataset was collected over the period of September 2021 to February 2022. The initial annotation guideline was developed while studying ELI5 and NQ answers, and the annotation on WebGPT answers [2] were collected in the last two months.

Does the dataset contain all possible instances? No. Our dataset provides annotated samples of QA pairs drawn from three existing LFQA datasets, covering a wide range of answers, including answers provided by users in online community (ELI5), answers written by trained annotators through web search (WebGPT), and answers identified in Wikipedia passages (NQ). Our data and analysis revealed different discourse structure in

[2]WebGPT answers were released in December, 2021.

different types of long-form answers. However, we acknowledge that our data as well as ontology does not cover all possible types of long-form answers, such as those derived from textbooks.

If the dataset is a sample, then what is the population? Our dataset represents a subset of information-seeking questions requiring a long-form answers. It does not cover the entire range of such (question, answer) pairs, as the datasets we source questions from have their own preprocessing methods. We also only consider answerable questions in NQ, while many of the unanswerable questions are of a similar nature but not considered as no answers exist in a single Wikipedia page. Our dataset also only covers questions and answer written in English.

4 Data Preprocessing

What preprocessing / cleaning was done? We first preprocess NQ to derive a filtered set of complex questions and then perform preprocessing on all LFQA datasets considered. We describe each step below.

Natural Question While WebGPT and ELI5 are question answering datasets only with long-form answers, NQ is normally studied under the setting of extractive / short-form QA. We thus create a filtered set of NQ that focuses on complex queries requiring long-form answers in the format of paragraph (i.e. excluding tables). We build a classifier, which selects 3,910 NQ questions (which we release as **NQ-complex** in our repository), roughly 10% of the 27,752 NQ examples with only long form answers from both the training and validation splits of the original NQ data.

We describe this classifier below, which was trained to distinguish NQ questions with only short answers and ELI5 questions from the question text alone. We build a simple BERT-based classifier, trained to distinguish NQ questions with short answers (i.e., less than five tokens) and ELI5 questions. We use the `[CLS]` token from BERT model to perform prediction. We use the original split from the ELI5 dataset and remove the questions whose answer's length is less than 3 sentences or longer than 15 sentences, resulting in 157,926, 5,695 and 14,186 questions in training, validation and test set. We use the training set of NQ questions with short answer for training and split the validation set for evaluation resulting in 35,989, 5,695 and 1,410 in the training, validation and test set. We preprocessed the questions by converting to lowercase and exclude punctuation to remove syntactic differences between ELI5 and NQ questions.

We fine-tuned the `bert-base-uncased` model for 3 epochs, with an initial learning rate of $5e-5$ and batch size of 32. We use the model with the highest validation F1 as the question classifier, which achieves F1 of 0.97 and 0.94 on validation and test set respectively. We then run this classifier to select the complex questions from NQ questions with long-form answers.

All After identifying the NQ complex questions, we preprocess all long-form QA data from ELI5, NQ-complex, WebGPT and ELI5-model by removing answers with more than 15 sentences and those with less than 3 sentences to make annotation task more manageable. We used Stanza (Qi et al., 2020) to split long-form answers into sentences. This process removes 42%, 28% and 34% from ELI5, WebGPT and NQ-complex respectively. We then randomly sample (question, answer) pairs from each dataset to conduct annotation.

Was the raw data saved in addition to the cleaned data? For the purpose of studying discourse structure, we only include the preprocessed data with either validity or role annotations. However, we released all NQ-complex questions identified, without filtering based on answer length.

Does this dataset collection/preprocessing procedure achieve the initial motivation? Our preprocessing on NQ questions identifies complex questions that require a paragraph-level answers and allows us to study long-form answers pre-existing in Wikipedia passages, which provides a complementary view compared to the other two datasets. While our preprocessing removed answers that are either too short or too long, we are able to observe consistent and interesting trend in the long-form answers we investigated.

5 Dataset Distribution

How is the dataset distributed? All annotated data is available at `https://github.com/utcsnlp/lfqa_discourse`.

When was it released? March 2022.

What license (if any) is it distributed under? The data is distributed under the CC BY-SA 4.0 license.[3]

Who is supporting and maintaining the dataset? The dataset will be maintained by the authors of this paper. Upates will be posted at https://github.com/utcsnlp/lfqa_discourse.

6 Legal and Ethical Considerations

Were workers told what the dataset would be used for and did they consent? Both crowdworkers and undergraduate annotators were informed of the goals of our study: to better understand what composes an answer to complex queries. Crowdworker consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement, while undergraduate annotators consented through hiring process of our institution.

If it relates to people, could this dataset expose people to harm or legal action? Our dataset does not contain any personal information of our annotators. However, our dataset contains samples sourced from existing, publicly available long-form question answering datasets which might contain incorrect and outdated information, and should be used with caution for such purpose.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? Our datasets are mainly sourced from English-speaking users and hence it might reflect societal biases and overlook culture and community whose primary language is non-English. We hope that future research could look into analyzing long-form answers in other languages.

References

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Fangyuan Xu, Junyi Jessy Li, and Eunsol Choi. 2022. How do we answer complex questions: Discourse structure of long-form answers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Long paper.

[3] https://creativecommons.org/licenses/by-sa/4.0/legalcode