Datasheet for SITUATEDQA

Michael J.Q. Zhang and **Eunsol Choi**
The University of Texas at Austin
{mjqzhang, eunsol}@utexas.edu

1 Motivation for Datasheet Creation

Why was the dataset created? Answers to the same question may change depending on the extra-linguistic contexts (when and where the question was asked). To study this challenge, we introduce SITUATEDQA (Zhang and Choi, 2021), an open-retrieval QA dataset where systems must produce the correct answer to a question given the temporal or geographical context. The standard paradigm for evaluating QA systems makes implicit assumptions about time and location and does not measure how well QA models adapt to new contexts. Therefore, our goals when collecting this SITUATEDQAwere to (1) evaluate open-retrieval QA systems on how they perform in these more practical settings and (2) facilitate the training and development of models that explicitly model how answers change across different extra-linguistic contexts.

Has the dataset been used already? All papers reporting on SITUATEDQA must submit their results to https://situatedqa.github.io.

Who funded the dataset? SITUATEDQA was funded by Google Faculty Awards and by UT Austin.

2 Dataset Composition

What are the instances? Each instance is an information seeking question that is annotated for its temporal or geographical dependence (i.e., whether the answer to the question depends on when or where it was asked). For some context dependent questions, we also include annotations of how the answer changes across different temporal or geographical contexts.

How many instances are there? Our dataset consists of 8.9K questions from one of from four existing datasets (Kwiatkowski et al., 2019; Berant et al., 2013; Clark et al., 2020; Campos et al., 2016) that are annotated with their temporal dependence and 2.4K questions for their geographical dependence. For 2.8K of those context-dependent questions from NQ-Open (2.4K temporal, 0.5K geographical), we collected a total of 5.9K answers from alternate contexts (4.0K temporal, 1.9K geographical). From those alternate temporal context / answer pairs, we construct (q, c_v, a) by sampling valid dates, creating 6K examples. The final TEMP dataset also includes 6.7K examples from temporally-independent questions. The fine-grained breakdown of the number of examples that have been identified as temporally or geographically dependent is included in Table 1. We also include the number of question, context, and answer triples for each context type in Table 2.

Context Type (c_t)	Train	Dev	Test
TEMP	4438	2572	1962
GEO	1149	879	367

Table 1: Breakdown of collected examples for context dependent question identification.

Context Type (c_t)	Train	Dev	Test
TEMP	6009	3423	2795
GEO	3548	1398	506

Table 2: Breakdown of collected examples for context dependent question answering.

What data does each instance consist of? Each example in SITUATEDQA consists of a question q, context c_i, and answer a_i where a_i is the answer to q when situated in the context c_i. Each context consists of a type c_{t_i} and a value c_{v_i}. We study two context types: temporal (TEMP) and geographical (GEO). TEMP defines each context value as timestamp (e.g. a date or year) where a_i is the answer to q if it was asked at the time of c_{v_i}. GEO defines each context value as a geopolitical entity where

Question q	Context Type c_t	Context Value c_v	Answer a
Who composed the music for the first Harry Potter film?	-	-	-
What's the biggest country in Europe excluding Russia?	-	-	-
How many seasons are there for American Horror Story?	TEMP	Sep 18, 2019 Sep 13, 2017	10 9
Who made the most three point shots in the NBA?	TEMP	2014 2005	Ray Allen Reggie Miller
When was the last time states were created?	GEO	Nigeria United States	1 October 1996 1959
Where do we rank among the world's largest cities?	GEO	Tokyo Shanghai	1st 3rd

Table 3: Examples of how questions interact with geographical and temporal context in SITUATEDQA. The first two questions are not identified as geographically nor temporally dependent.

a_i is the answer to the q in the location c_{v_i}. See Table 3 for examples from each context type.

For examples with only context-dependent question identification labels, each example consists of a question q, context type c_t (TEMP or GEO), as well as a binary label determining whether there exists two distinct contexts values, (c_{v_i}, c_{v_j}), with different respective answers, $a_i \neq a_j$.

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

Are there recommended data splits or evaluation measures? We include the recommended train, development, and test sets for our datasets. For context dependent QA examples, we also include "easy" and "hard" subsets of our evaluation data, determined by what contexts require models to generalize to new contexts. We use standard evaluation measures for context dependent QI: binary classification accuracy, precision, recall, and F1. For context dependent QA, we use exact match accuracy following the answer normalization steps from Chen et al. (2017).

3 Data Collection Process

How was the data collected? We split up data collection into three stages: (1) Identification where annotators are asked to identify whether to the answer to a question depends on its temporal or geographical context. (2) {Context / Answer} collection where annotators are asked to identify the answers from additional contexts by providing a brief timeline of answers to temporally dependent questions or valid location/answer pairs for geographically dependent questions. (3) Validation where annotators are presented with all answer timelines or location/answer pairs that were collected in the prior step and are asked to validate or revise each.

Who was involved in the collection process and what were their roles? We recruit crowdworkers from Amazon Mechanical Turk to perform the all the annotation steps outlined above.

Over what time frame was the data collected? The dataset was collected over a period of October 2020 to March 2021. Annotations from the {Context / Answer} collection and validation steps (described above) was collected in the last two months.

Does the dataset contain all possible instances? We source our questions from a variety of existing datasets for open retrieval QA; however, none of these datasets cover the full range of information seeking questions. For instance, many ambiguous questions are filtered out when constructing open retrieval QA datasets due to their ambiguity. These filtering methods often remove many naturally occurring geographically dependent questions. Our annotations also do not cover the full range of possible contexts. While our dataset is aimed at covering a wide range of possible temporal contexts and locations, we still only cover a fraction of the space of all possible contexts. These contexts are also biased by what information is available and is easily accessible on Wikipedia. Our annotation process, however, encourages workers to find the most relevant temporally deponent answers (the two most recent) and geographically dependent answers from a wide range of possible locations. Having workers retrieve answers from multiple locations encourages then to seek out rare locations that are poorly represented in existing datasets.

If the dataset is a sample, then what is the population? Our dataset represents a subset of information seeking questions and their answers from different temporal and geographical contexts. It does not cover the entire range of information seeking questions, as the datasets we source questions from have their own sampling methods. Furthermore, all questions in our dataset are answerable by Wikipedia documents; however, there are many facts that are only available in other sources such as news articles. It also only covers two most recent answers (at the time of collection) to temporally dependent questions, and answers from a few of the vast range of possible geographical contexts. Our dataset also only covers questions and answers written in English.

4 Data Preprocessing

What preprocessing / cleaning was done? Our questions are sampled from a variety of open-retrieval QA datasets (Kwiatkowski et al., 2019; Berant et al., 2013; Clark et al., 2020; Campos et al., 2016). To generate geographically dependent questions by modifying existing questions. We identify questions with phrases that specify a location by running an NER tagger (Peters et al., 2017) and remove the location entity using heuristics based on its syntactic role as identified by a dependency parser (Dozat and Manning, 2017). If the entity's syntactic role is either *nn* or *amod*, we delete the entire entity and all of its descendants. If the the entity's role is *pobj*, the entity's parent preposition and all its descendants. Finally, if the entity's role is *root* or *nsubj*, we replace the entity with the pronoun *we*, deleting all determiners and conjugating any auxiliary verbs accordingly. We ignore instances where the dependency is not in one of these categories, there are multiple GPE entities, the stripped questions is has 3 tokens or less, or there is disagreement between our parser and tagger. We use the implementations of Peters et al. (2017) and Dozat and Manning (2017) from AllenNLP (Gardner et al., 2017)

We construct TEMP examples in one of three ways: (1) We use each answer's start transition timestamp as c_v. (2) We uniformly sample up to two dates/years between each answer's start and end transitions, using each sampled timestamp to create a new (q, c_v, a) triple. (3) We use questions that were annotated as **not** temporally dependent by, we uniformly sample a single value of c_v between 2018 and March 2021, resulting in one (q, c_v, a) triple per static question.

Was the raw data saved in addition to the cleaned data? We include the raw answer timelines and annotated location/answer pairs after validation. We also release the annotations from the identification stage, including the examples that were later filtered out.

Does this dataset collection/preprocessing procedure achieve the initial motivation? Our collection process indeed achieves our initial goals of creating a dataset for exploring the role of extra-linguistic contexts in information seeking question answering. Using this data, we are able to evaluate how models that are trained on past data generalize to answering questions in the future, asked at the time of our data collection. We recognize, however, that as time goes on, facts will continue to change and will necessitate updating benchmarks with new answers.

5 Dataset Distribution

How is the dataset distributed? SITUATEDQA is availible at `https://situatedqa.github.io`.

When was it released? September 2021.

What license (if any) is it distributed under? SITUATEDQA is distributed under the CC BY-SA 4.0 license.[1]

Who is supporting and maintaining the dataset? This dataset will be maintained by the authors of this paper. Updates will be posted at `https://situatedqa.github.io`.

6 Legal and Ethical Considerations

Were workers told what the dataset would be used for and did they consent? Crowd workers informed of the goals we sought to achieve through data collection: to improve the ability of QA systems to handle different extra-linguistic contexts. They consented to have their responses used in this way through the Amazon Mechanical Turk Participation Agreement.

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

If it relates to people, could this dataset expose people to harm or legal action? Our dataset does not contain any personal information of crowd workers; however, our dataset can include incorrect information. We perform extensive quality control and error analysis to minimize the risk due to incorrect facts.

If it relates to people, does it unfairly advantage or disadvantage a particular social group? One of our goals in collecting this dataset was to facilitate the development of QA systems that can answer questions from people who live in locations that are poorly represented in current datasets. While our dataset does cover many of these poorly represented locations, we note that it only covers a fraction of the long tail of possible locations. Furthermore, there may be bias in what locations are covered, in part due to what information is available and easily accessible on Wikipedia.

References

Jonathan Berant, Andrew K. Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*.

Daniel Fernando Campos, T. Nguyen, M. Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, L. Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

J. Clark, Eunsol Choi, M. Collins, Dan Garrette, T. Kwiatkowski, V. Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. *ArXiv*, abs/1611.01734.

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.

T. Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, C. Alberti, D. Epstein, Illia Polosukhin, J. 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*, 7:453–466.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and R. Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *ACL*.

Michael J.Q. Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. *EMNLP*.