A Appendix for "A Dataset for Answering Time-Sensitive Questions"

A.1 Dataset documentation and intended uses

We follow datasheets for datasets guideline to document the followings.

A.1.1 Motivation

- For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?
  TimeQA is created to test current models' capability to perform diverse temporal reasoning under unstructured text corpus, which can help future NLP models to better capture the time dimension.
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
  UCSB NLP team, mostly Wenhu Chen

A.1.2 Composition

- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?
  TimeQA only contains documents (text) in the dataset.
- How many instances are there in total (of each type, if appropriate)?
  There are roughly 20K question-answer pairs.
- Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
  It's sampled from large Wikipedia passages, it's representative of all the possible temporal-sensitive information.
- Are relationships between individual instances made explicitly (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
  N/A.
- Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
  Yes, we split training, development, and testing set. We split randomly within each data source.
- Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
  There could have some potential noise of question or answer annotation.
- Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions] (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
  TimeQA is self-contained.
- Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
  No, all the samples in TimeQA is public available.

- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
  No
- Does the dataset relate to people? If not, you may skip the remaining questions in this section.
  No

A.1.3 Uses

- Has the dataset been used for any tasks already? If so, please provide a description?
  It is proposed to use for QA task.
- Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
  It is a new dataset. We run existing state-of-the-art models and release the code at https: https://github.com/wenhuchen/Time-Sensitive-QA
- What (other) tasks could the dataset be used for?
  Many other tasks like relation extractions can be also used.
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
  N/A
- Are there tasks for which the dataset should not be used? If so, please provide a description.
  N/A

A.1.4 Distribution

- Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
  No.
- How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
  Release on Github. No DOI
- When will the dataset be distributed?
  It is released in https://github.com/wenhuchen/Time-Sensitive-QA
- Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
  BSD 3-Clause "New" or "Revised" License. https://github.com/wenhuchen/Time-Sensitive-QA/blob/main/LICENSE
- Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
  No.
- Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

A.1.5 Accessibility

- Links to access the dataset and its metadata: the github repository https://github.com/wenhuchen/Time-Sensitive-QA.
- The data is saved in a json format, where an example is shown in the README.md file.

- UCSB NLP group will maintain this dataset on the Github account.
- BSD 3-Clause "New" or "Revised" License https://github.com/wenhuchen/Time-Sensitive-QA/blob/main/LICENSE

A.2 Evaluation Metrics

The F1 score is calculated with the following equation to cover the '[unanswerable]' case.

$$F1(\mathbb{A}_j^{(i)}, \hat{\mathbb{A}}^{(i)}) = \begin{cases} 1, & if A_j^{(i)} = NULL \ \& \ \hat{A}^{(i)} = NULL \\ 0, & if A_j^{(i)} = NULL \ \& \ \hat{A}^{(i)} \neq NULL \\ 0, & if A_j^{(i)} \neq NULL \ \& \ \hat{A}^{(i)} = NULL \\ f1(\mathbb{A}_j^{(i)}, \hat{\mathbb{A}}^{(i)}), & if A_j^{(i)} \neq NULL \ \& \ \hat{A}^{(i)} \neq NULL \end{cases}$$

A.3 Annotation Interface

The annotation interface is demonstrated in Figure 8, the original HIT job url is `https://s3.amazonaws.com/mturk_bulk/hits/467870457/JV_fEwGzVy_PgHqtWqWXLA.html`.

Figure 8: The annotation interface for crowd workers.

A.4 Relation Distribution over TimeQA

The annotated time-evolving facts include more than 70 relations, which follows a long-tail distribution. The major relations are shown in Figure 9, where the dominant relations are 'P54' (play for), 'P39' (position held), 'P108' (employer of), 'P69' (educated at).

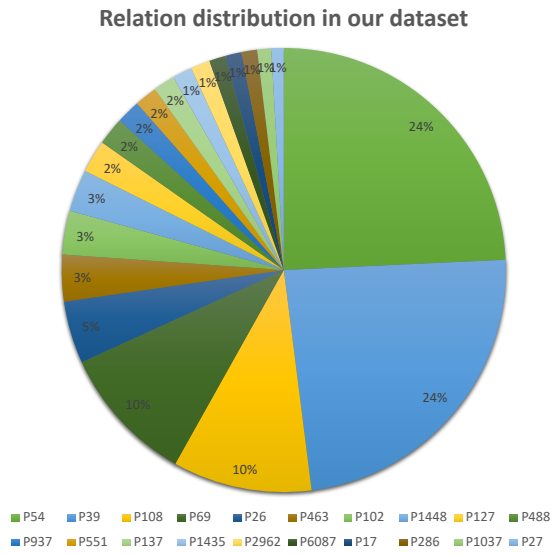

Figure 9: The relation distribution over the annotated facts.

A.5 Answerable vs. Unanswerable

We also provide break-down analysis of model performance for answerable and unanswerable questions in Figure 10. As can be seen, the FiD is more aware of the answerability on TimeQA. On the easy and hard mode, FiD's accuracy on unanswerable is above BigBird's by 14% and 18%, while the gap in answerable questions are less significant.

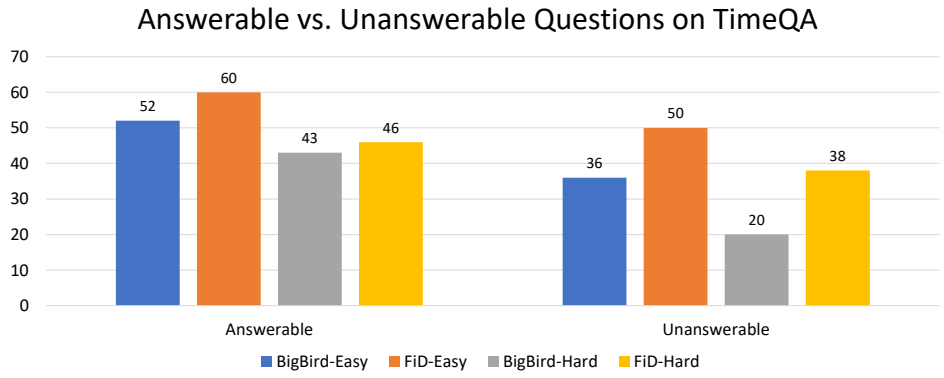

Figure 10: The answerable and unanswerable question performance for BigBird and FiD.