Supplementary File for ExpMRC: Explainability Evaluation for Machine Reading Comprehension

Yiming Cui[1,2], **Ting Liu**[1], **Wanxiang Che**[1], **Zhigang Chen**[2], **Shijin Wang**[2,3]
[1]Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, China
[3]iFLYTEK AI Research (Hebei), Langfang, China
[1]{ymcui,tliu,car}@ir.hit.edu.cn
[2,3]{ymcui,zgchen,sjwang3}@iflytek.com

1 Dataset Documentation

In this section, we elaborate all details for the key requirements in CFP of NeurIPS 2021 Dataset Track.

1.1 Datasheets for Datasets

Following Gebru et al. (2020), we describe our dataset as follows, including motivations, composition, collect process, preprocessing/cleaning/labeling, uses, distribution, and maintenance.

1.1.1 Motivations

Q1: For what purpose was the dataset created?

The proposed ExpMRC is built to accelerate the research of the explainability on MRC models. We believe the inclusion of our dataset could enrich the explanatory studies in NLP.

Q2: Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset is created by the reading comprehension group of the Joint Laboratory of HIT and iFLYTEK Research.

Q3: Who funded the creation of the dataset?

Yiming Cui is partially supported by Google TPU Research Cloud (TRC) program. This work was supported by the National Key R&D Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 61976072 and 61772153.

1.1.2 Composition

Q1: What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

Documents. Specifically, Wikipedia passages and examinations for middle/high school students.

Q2: How many instances are there in total (of each type, if appropriate)?

The statistics are stated in Table 2 of the main text, where we also post here.

Q3: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Table 1: Statistics of the proposed ExpMRC.

	SQuAD		CMRC 2018		RACE$^+$		C^3	
	Dev	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Dev**	**Test**
Language	English		Chinese		English		Chinese	
Answer Type	passage span		passage span		multi-choice		multi-choice	
Domain	Wikipedia		Wikipedia		exams		exams	
Passage #	319	313	369	399	167	168	273	244
Question #	501	502	515	500	561	564	505	500
Max Answer #	3	3	3	3	1	1	1	1
Max Evidence #	2	2	3	3	2	2	4	4
Avg/Max P #	146/369	157/352	467/961	468/930	311/514	324/603	426/1096	413/1011
Avg/Max Q #	12/28	11/28	15/37	15/37	15/39	16/55	14/28	14/31
Avg/Max A #	3/25	3/27	6/64	5/33	6/20	6/27	7/25	7/35
Avg/Max E #	26/62	28/76	43/175	52/313	23/162	23/82	37/199	41/180

[Yes] Only the samples that meet our annotation criteria are kept, as illustrated in Section 3.2 of the main paper.

Q4: What data does each instance consist of?

Each instance consists of raw passage, question, candidate (if applicable), answer, and evidence sentence, with several meta-data (such as question_id).

Q5: Is there a label or target associated with each instance?

[Yes] The ground truth answer and evidence is provided in the dataset.

Q6: Is any information missing from individual instances?

[No]

Q7: Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Each instances are independent.

Q8: Are there recommended data splits (e.g., training, development/validation, testing)?

[Yes] Please refer to Table 1.

Q9: Are there any errors, sources of noise, or redundancies in the dataset?

[No] The errors and noises are eliminated during the annotation process.

Q10: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

[Yes]

Q11: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)?

[No]

Q12: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

[No] Such data is not included in our dataset.

Q13: Does the dataset relate to people?

[Yes]

Q14: Does the dataset identify any subpopulations (e.g., by age, gender)?

[No]

Q15: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

[No]

Q16: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

[No]

1.1.3 Collection Process

Q1: How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?

The data is directly observable, as all data is in text form.

Q2: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, softwareAPI)?

We use in-house implemented web platform for the annotation process. It is used for various dataset annotation in our institute.

Q3: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Only the samples that meet our annotation criteria are kept, as illustrated in Section 3.2 of the main paper.

Q4: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

We have illustrated in Section 3.2 of the main paper. The annotators are either English-majored or Chinese-majored graduate students from China, depending on the dataset language. All annotators are full-intern students, and are paid monthly with proper internship salaries (approximately \$400 to \$500). Each evidence piece costs approximately \$0.50 for all types of MRC data. \$0.50 per evidence is an internal price for managing the project, and estimate the internal cost of the whole project.

Q5: Over what timeframe was the data collected?

For SQuAD, CMRC 2018, and C^3, they are already publicly available. For $RACE^+$, it was collected during year 2018. The annotation of the evidence is performed during Nov 2020 to Jan 2021.

Q6: Were any ethical review processes conducted (e.g., by an institutional review board)?

[No] But we have checked these data by ourselves to ensure there is no ethical issues.

1.1.4 Preprocessing, Cleaning, Labeling

Q1: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

[Yes]

Q2: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

[No]

Q3: Is the software used to preprocess/clean/label the instances available?

[No]

1.1.5 Uses

Q1: Has the dataset been used for any tasks already?

[No]

Q2: Is there a repository that links to any or all papers or systems that use the dataset?

[No] Our paper has just made public on arXiv, and it has no citations.

Q3: What (other) tasks could the dataset be used for?

The proposed ExpMRC is mainly designed for Explainable MRC. It can be used for evaluating the quality of the answer as well as its explanations. Also it can be used to perform analyses on how the model solves the questions. Moreover, it can be used with other datasets to discover different explainable behavior inside the model.

Q4: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

[No]

Q5: Are there tasks for which the dataset should not be used?

[No]

1.1.6 Distribution

Q1: 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?

[No]

Q2: How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?

The dataset will be distributed on GitHub:

```
https://github.com/ymcui/expmrc
```

Q3: When will the dataset be distributed?

The dataset has already been released on Github.

Q4: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset is distributed under CC BY-SA 4.0 license.

Q5: Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

[No]

Q6: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

[No]

1.1.7 Maintenance

Q1: Who is supporting/hosting/maintaining the dataset?

Yiming Cui and his team are in charge of maintaining the dataset.

Q2: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The participants could contact one of the authors, preferably to Yiming Cui (ymcui@ir.hit.edu.cn) or our mailing group (expmrc@126.com).

Q3: Is there an erratum?

[No]

Q4: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

[No] But we will update our dataset if there is a strong request from our community. Such modifications will be made public via our GitHub repository.

Q5: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

[No]

Q6: Will older versions of the dataset continue to be supported/hosted/maintained?

[Yes] If new versions become available, we will also maintain the old versions for a period till a complete transition to the new version.

Q7: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

[Yes] As long as they observe the CC-BY-SA 4.0 license.

1.2 Dataset URL

Our dataset and baseline codes are available:

`https://github.com/ymcui/expmrc`

1.3 Hosting, Licensing, and Maintenance Plan

We have set our leaderboard at `https://ymcui.com/expmrc`, which is based on the GitHub Pages. This ensures our website will be hosted properly.

Our testing server is based on CodaLab Worksheet[1], which is an open-source platform for reproducible papers. We have set up proper submission guidelines:

`https://worksheets.codalab.org/worksheets/0xfe1ce37d9d2e45aa927a78289c548489`

Figure 1 shows a screenshot of our submission site.

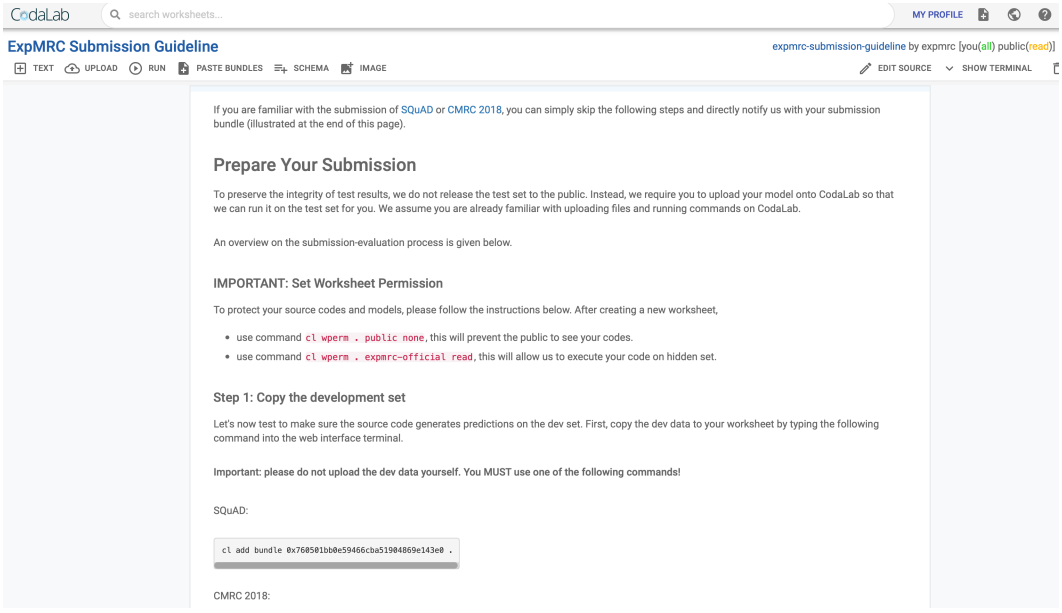

Figure 1: Submission site of ExpMRC.

[1] `http://worksheets.codalab.org`

Our dataset is distributed under CC BY-SA 4.0 license[2] and the baseline codes are distributed under Apache-2.0 license[3]. The first author has been hosting the *Workshop on Chinese Machine Reading Comprehension (CMRC)* for many years, and we are still accepting CMRC 2017[4], CMRC 2018[5], CMRC 2019[6] submissions to allow further test on the hidden test set. In this context, we will also maintain our ExpMRC leaderboard for a long time, as the explainability in ML is one of the most trending topic in recent research.

2 Comments from Previous Venue

Our paper was previously submitted to ACL 2021 and received a borderline score of 3.33 (out of 5), which was possibly accepted to Finding of ACL 2021 (a companion publication to ACL 2021), but was not selected to publish. While some of the concerns were resolved during the author response phase in previous venue, due to the limit text space for the rebuttal, we were unable to provide point-to-point reply to every concerns.

Overall, the reviewers agree that our ExpMRC is well-motivated and this is a timely work for explainable AI. However, they also have some concerns, where we list them as follows.

- Lack of detailed illustration on the annotation procedures.
- The selection of the lambda term may not be optimal, and the experimental results as well as the analysis should be updated.
- There are some grammatical issues and some of the content needs revision for better presentation.

In this version, we have resolved the concerns by the reviewers, where we briefly conclude as follows.

- Within the paper length limit, we have enriched the illustrations on the data annotation process. We also present a detailed documentation in this supplementary file.
- We have tested the lambda term under 0.1 and found that lambda=0.01 seems to be the optimal value for span-extraction MRC tasks (SQuAD and CMRC 2018). We have updated experimental results and the analysis part to reflect our new results.
- We have double checked our manuscript and also get our paper proofread by a professional native speaker.

References

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. 2020. Datasheets for Datasets. arXiv:cs.DB/1803.09010

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

[3] https://www.apache.org/licenses/LICENSE-2.0

[4] https://github.com/ymcui/cmrc2017

[5] https://github.com/ymcui/cmrc2018

[6] https://github.com/ymcui/cmrc2019