Datasheet for RecipeQA*

1 Motivation for Dataset Creation

Why was the dataset created?

RecipeQA was created to facilitate research on multimodal machine comprehension which involves understanding procedural data given in the form of cooking recipes with accompanying images.

Has the dataset been used for any tasks already?

All papers reporting results on RecipeQA are required to submit their results to the project webpage at http://hucvl.github.io/recipeqa.

Who funded the creation of the dataset?

RecipeQA was supported in part by a Hacettepe BAP fellowship (FBB-2016-11653) awarded to Erkut Erdem.

2 Dataset Composition

What are the instances?

Each instance in RecipeQA includes a context passage in the form of a recipe, a question denoting a specific comprehension task, and four candidate answers. For the question being asked there is only a single correct answer. At present, there are four tasks in RecipeQA, namely *textual cloze*, *visual cloze*, *visual ordering* and *visual coherence*. Depending on the task, the context passage might contain, in addition to step-by-step textual instructions, a number of illustrative images and the questions and answers might consist of images or text.

How many instances are there?

RecipeQA consists of 36,786 question-answer pairs which are generated automatically from approximately 19,779 unique cooking recipes. A total of 3567 question-answer pairs are kept private for evaluation purposes.

What data does each instance consist of?

The instances come from the cooking recipes collected from instructibles.com. Each recipe includes an arbitrary number of steps containing both textual and visual elements. In particular, each step of a recipe is accompanied by a 'title', a 'description' and a set of illustrative 'images' that are aligned with the title and the description. Each of these elements can be considered as a different modality of the data. The questions in RecipeQA explore the multimodal aspects of the step-by-step instructions available in the recipes through a number of specific tasks.

Does the data rely on external resources?

Everything is included in the data release.

Are there recommended data splits and evaluation measures?

The release comes with non-overlapping train, validation and test splits with only the training and validation sets being released publicly. We keep the test set private and encourage the researchers to submit their models to the challenge website at hucvl.github.io/recipeqa, which evaluates and reports the submitted models on our hidden test set. We also provide an official evaluation script for RecipeQA used by our leaderboard site for evaluation.

3 Data Collection Process

How was the data collected?

We consider cooking recipes as the main data source for our dataset. These recipes were collected from instructables.com, which is a how-to web site where the users share all kinds of instructions including but not limited to cooking recipes.

We employed a set of heuristics that helped us collect high quality data in an automatic manner. For instance, while collecting the recipes, we downloaded only the most popular recipes by considering the popularity as an objective measure for assessing the quality of a recipe.

*Prepared in accordance with the guideline suggested in (Gebru et al., 2018).

Who was involved in the data collection process?

We (the authors) did initial analyses on the collected data. Finally, we gathered the recipes only with the desired attributes and information.

Over what time-frame was the data collected?

All materials from the instructables.com were downloaded in April 2018 over a month period.

Does the dataset contain all possible instances? No, we release the training and validation sets and keep the test set private for evaluation purposes. It is also possible to generate new question-answer pairs using the recipes found in the dataset.

If the dataset is a sample, then what is the population?

The recipes released in the RecipeQA are not representative of the all cooking recipes found on instructables.com or other cooking recipe websites. We downloaded only the most popular recipes by considering the popularity as an objective measure for assessing the quality of a recipe. Our assumption is that the mostly viewed recipes contain less noise and include easy-to-understand instructions with high-quality illustrative images.

4 Data Preprocessing

What preprocessing/cleaning was done?

We filtered out non-English recipes using a language identification tool (Lui and Baldwin, 2012), and automatically removed the ones with unreadable contents such as the ones that only contain recipe videos. Finally, as a post processing step, we normalized the description text by removing non-ASCII characters from the text.

In order to generate question-answer-context triplets, we first filtered out recipes that contain less than 3 steps or more than 25 steps. We also ignored the initial step of the recipes as our preliminary analysis showed that the first step of the recipes almost always is used by the authors to provide a narrative.

Was the "raw" data saved in addition to the preprocessed/cleaned data?

The raw unprocessed data is saved but it did not make public, we only released the instances that we cleaned up and preprocessed.

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

RecipeQA indeed serves as a challenging test bed and an ideal benchmark for evaluating machine comprehension systems. It will especially facilitate research on both procedural information and multimodal comprehension problems.

5 Dataset Distribution

How is the dataset be distributed?

The dataset is available at the project webpage at hucvl.github.io/recipeqa.

When was it released? September 2018

What license (if any) is it distributed under?

RecipeQA contains QA pairs generated from the cooking recipes which are shared publicly with a variety of licences. Corresponding licence for each recipe is provided in the dataset, see `recipes.json`. Additionally, the researchers that use RecipeQA are requested to cite the corresponding dataset paper.

6 Dataset Maintenance

Who is supporting and maintaining the dataset?

The dataset will be maintained by the authors of the paper: Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. All updates will be posted on the dataset website.

Will the dataset be updated? Extending the dataset with new recipes, tasks and question-answer pairs is planned. All changes to the dataset will be announced through the dataset and the challenge website at http://hucvl.github.io/recipeqa.

If the dataset becomes obsolete how will this be communicated?

This will be posted on the dataset webpage.

7 Legal & Ethical Considerations

Were workers told what the dataset would be used for and did they consent?

The cooking recipes in RecipeQA were gathered from the how-to documents related to the food category on instructibles.com. All the recipes have the Creative Commons licenses which lets others use and modify the provided content. Hence, the recipes in the RecipeQA already have proper consent from the authors of the recipes.

If it relates to people, could this dataset expose people to harm or legal action?

No, the collected cooking recipes were already public.

If it relates to people, does it unfairly advantage or disadvantage a particular social group?

The cooking recipes selected for RecipeQA might have introduced some biases as they reflect tastes and preferences of instructibles.com users. Moreover, our preprocessing involves filtering out non-English recipes and therefore the collected recipes represent only a fraction of foods from around the world.

References

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Marco Lui and Timothy Baldwin. 2012. `langid.py`: An off-the-shelf language identification tool. In *Association for Computational Linguistics (ACL) Demo Session*, pages 25–30.