Completed Worksheet: BoolQ
Evidence-Centered Benchmark Design

Evidence-Centered Benchmark Design (ECBD) is a framework that formalizes the benchmark design process. It requires first specifying the **intended use** of the benchmark (including specifying the objects of evaluation). The process is then broken down into five modules:

- **Capability module:** capabilities that the benchmark aims to measure.
- **Content module:** pool of test items that draw out responses from the objects.
- **Adaptation module:** adapting or instructing the objects to complete the tasks.
- **Assembly module:** selecting from the pool of test items to build the set used for evaluation.
- **Evidence module:** extracting and accumulating evidence about the capabilities of interest from responses produced by the objects.

This worksheet provides guidance on how to create a new benchmark or analyze an existing benchmark following ECBD. It can be completed from different perspectives: as the creator of a new benchmark, as the custodian or the user of an existing benchmark, or as a third-party analyzing benchmarks, etc. Each module contains three questions:

- **Describe:** what design decisions did the benchmark creators make for this module?
- **Justify:** why did the benchmark creators make these decisions? This involves forming a hypothesis that the decisions allow the module to accomplish its role in the process of gathering necessary capability evidence.
- **Evidence:** what validity evidence do the benchmark creators have to support the above hypothesis? In other words, what shows that the module indeed accomplishes its role?

This worksheet is not a checklist, and it is not required to answer each question perfectly. These questions are meant to encourage reflection and validation of benchmark design decisions, as well as to guide benchmark documentation.

Table of Contents

Benchmark Name and Reference(s)

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In North American Chapter of the Association for Computational Linguistics (NAACL).

URL: https://aclanthology.org/N19-1300/

Who is filing the worksheet:

From what perspective is this worksheet completed? In other words, what is the relation between the person(s) completing this worksheet and the benchmark that is the focus of this worksheet?

Third-party: Yu Lu Liu, Jackie Chi Kit Cheung, Alexandra Olteanu

Intended Use

The validity of a benchmark concerns whether it can be used as intended. It is therefore crucial to first clearly establish the intended use of a benchmark before analyzing it or creating it.

Q1 - Who/What are the intended objects of evaluation? Elaboration on the objects of evaluation (e.g., their assumed capabilities, demographic information for human objects of evaluation, etc.) helps us better understand whether the benchmark is suitable for all intended objects of evaluation.

The targetted evaluatees are **"models"** (p.2924), which includes (according to the models evaluated in their experiments) shallow models as well as neural models such as BERT, Elmo, GPT.

The authors also claim to have evaluated the ability of **humans** to answer the boolean questions in the benchmark.

Q2 - What is the intended use of the benchmark? Who are the intended users of the benchmark? Benchmark results aim to provide insights about the objects of evaluation: how are users meant to use these insights?

The authors do not explicitly mention any intended users of the benchmark, perhaps implied to be the NLP research community. The benchmark results would provide insights about the capability of interest (see Capability Module) of the evaluatees, but it is unstated what intended users would to do with these insights.

Based on the uses in the experiments in the paper, these results could be used to:

- Compare models against humans in terms of the capability of interest.
- Compare transfer learning methods to find the most effective approach to obtain more capable models (in terms of the capability of interest).

The Capability Module

The capability module specifies the *capabilities* that the benchmark aims to evaluate. The term "capability" refers to a construct (e.g., quality criteria, skill, etc.) that the objects of evaluation are thought to exhibit or possess. Capabilities often cannot be directly observed or directly measured, thus requiring the benchmark to indirectly measure them by gathering necessary evidence about said capabilities.

Q3 - DESCRIBE: i) What are the capabilities of interest? ii) How is each one defined, and under what context is each one defined?

There seems to be different level of granularity for the target capability. At the highest level, the capability of interest is to "understanding what facts can be inferred to be true or false from text [which is] an essential part of natural language understanding." At a more granular level, it's the "ability to answer naturally occurring yes/no questions" (p.2924)

The connection between these two capabilities seem to be that naturally occurring yes/no questions would often query "non-factoid information, and that human annotators need to apply a wide range of inferential abilities when answering them." (p. 2925) → the ability to answer naturally occurring yes/no questions would thus capture the highest-level capability of reasoning/understanding.

Additional recommended questions to consider so to further clarify and contextualize the definitions (in benchmark analysis: as presented by the benchmark)

- How does the definition used by the benchmark differ from other existing definitions of this capability?

  Nothing mentioned.

- How does this capability differ from other similarly defined capabilities?

  The same higher-level capability of reasoning/understanding is captured differently through:
  - Ability to distinguish entailment: label statements as "entailment" or "contradiction" based on a given text. The authors argue that "Using naturally occurring yes/no questions ensures even greater independence between the questions and premise text, and ties our dataset to a clear end-task".
  - Ability to answer questions: multi-step reasoning (HotPotQA), conversational QA (e.g., CoQA). The main difference is the *naturalness*. The authors criticizes prior datasets "heavily [prompting] users", "class imbalance (80% 'yes' answers)", and "engineering data to be more difficult". "This risks resulting in models that do not have obvious end-use applications since they are optimized to perform in an artificial setting" (p.2925)

Q4 - JUSTIFY: How are the capabilities of interest connected to the intended use of the benchmark (specified in Q2)? Are the capabilities theoretically attainable by the objects

to be evaluated? Explain the interest in measuring the capabilities in Q3 and question whether it may be impossible for the objects of evaluation to have said capabilities.

This question is difficult to answer given that the intended use (Q2) is poorly specified. There is no discussion of theoretical attainability in the paper.

Q5 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice and definition of capabilities of interest?

The data analysis (see Content Module) consistutes evidence that naturally-occurring yes/no questions collected in this benchmark i) query non-factoid information and ii) require human annotators to apply a wide range of inferential abilities to answer them.

However, we consider this data analysis to be evidence that the gathered dataset actually satisfies conditions i) and ii) -- not evidence showing that satisfying conditions i) and ii) *in general* capture the highest-level capability of reasoning/understanding, which what we need here in this module. (?)

The Content Module

The content module specifies test items that the benchmark could require objects of evaluation to perform or to respond to. The items should elicit evidence about some capability of interest, so that said capability evidence can be later extracted from the responses and aggregated to produce a measurement of said capability.

Q6 - DESCRIBE:

- **Characterize the test items**. Most often, NLP evaluation relies on input data, so this step could involve describing the data that is available to the benchmark to use, how the data is obtained, etc.

Each content instance consists of a boolean question, a paragraph from a Wikipedia article, and the title of the article. The questions are selected from queries to the Google search engine. The work followed the data collection method used by Natural Questions (NQ) (Kwiatkowski et al., 2019):

> *Annotators label question/article pairs in a three-step process. First, they decide if the question is good, meaning it is comprehensible, unambiguous, and requesting factual information. This judgment is made before the annotator sees the Wikipedia page. Next, for good questions, annotators find a passage within the document that contains enough information to answer the question. Annotators can mark questions as "not answerable" if the Wikipedia article does not contain the requested information." (p. 2926)*

The authors presented further information on the topic of the questions, what kind of information the questions request, and the kinds of reasoning skills required for humans to answer the questions (p. 2927-2928)

Question Topic			
Category	Example	Percent	Yes%
Entertainment Media	Is You and I by Lady Gaga a cover?	22.0	65.9
Nature/Science	Are there blue whales in the Atlantic Ocean?	22.0	56.8
Sports	Has the US men's team ever won the World Cup?	11.0	54.5
Law/Government	Is there a seat belt law in New Hampshire?	10.0	70.0
History	Were submarines used in the American Civil War?	5.0	70.0
Fictional Events	Is the Incredible Hulk part of the avengers?	4.0	87.5
Other	Is GDP per capita same as per capita income?	26.0	65.4
Question Type			
Category	Example	Percent	Yes%
Definitional	Is thread seal tape the same as Teflon tape?	14.5	55.2
Existence	Is there any dollar bill higher than a 100?	14.5	69.0
Event Occurrence	Did the great fire of London destroy St. Paul's Cathedral?	11.5	73.9
Other General Fact	Is there such thing as a dominant eye?	29.5	62.7
Other Entity Fact	Is the Arch in St. Louis a national park?	30.0	63.3

Table 1: Question categorization of BoolQ. Question topics are shown in the top half and question types are shown in the bottom half.

Reasoning Types		Yes/No Question Answering Examples
Paraphrasing (38.7%) The passage explicitly asserts or refutes what is stated in the question.	**Q**: **P**: **A**:	Is Tim Brown in the Hall of Fame? Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame. Yes. ["inducted into" directly implies he is in Hall of Fame.]
By Example (11.8%) The passage provides an example or counter-example to what is asserted by the question.	**Q**: **P**: **A**:	Are there any nuclear power plants in Michigan? . . . three nuclear power plants supply Michigan with about 30% of its electricity. Yes. [Since there must be at least three.]
Factual Reasoning (8.5%) Answering the question requires using world-knowledge to connect what is stated in the passage to the question.	**Q**: **P**: **A**:	Was designated survivor filmed in the White House? The series is. . . filmed in Toronto, Ontario. No. [The White House is not located in Toronto.]
Implicit (8.5%) The passage mentions or describes entities in the question in way that would not make sense if the answer was not yes/no.	**Q**: **P**: **A**:	Is static pressure the same as atmospheric pressure? The aircraft designer's objective is to ensure the pressure in the aircraft's static pressure system is as close as possible to the atmospheric pressure. . . No. [It would not make sense to bring them "as close as possible" if those terms referred to the same thing.]
Missing Mention (6.6%) We can conclude the answer is yes or no because, if this was not the case, it would have been mentioned in the passage.	**Q**: **P**: **A**:	Did Bonnie Blair's daughter make the Olympic team? Blair and Cruikshank have two children: a son, Grant, and daughter, Blair.... Blair Cruikshank competed at the 2018 United States Olympic speed skating trials at the 500 meter distance. No. [The passage describes Blair Cruikshank's daughter's skating accomplishments, so it would have mentioned it if she had qualified.]
Other Inference (25.9%) The passage states a fact that can be used to infer whether the answer is true or false, and does not fall into any of the other categories.	**Q**: **P**: **A**:	Is the sea snake the most venomous snake? . . . the venom of the inland taipan, drop by drop, is the most toxic among all snakes No. [If inland taipan is the most venomous snake, the sea snake must not be.]

Table 2: Kinds of reasoning needed in the BoolQ dataset.

Note that this described dataset is NOT the one used to evaluate human annotators. In fact, a small subset of it is taken as responses (see Evidence Module) from human annotators.

- **Which capabilities of interest does each item aim to capture?** Each item can aim to capture one or several capabilities amongst those listed in Q3.

As there is only one capability of interest (at the most granular level, and at the highest level), all content instances must aim to measure it.

Q7 - JUSTIFY: How does each test item elicit evidence about its target capabilities? Justify via the item descriptions above.

The test times are shown to satisfy the following criteria:
i) Querying non-factoid information (Table 1 in Q6)
ii) Humans need to apply a wide range of inferential abilities (Table 2 in Q6) when answering the questions.
The benchmark creators seem to be making the justification that those questions will, as a result, require models to demonstrate inferential abilities.

The benchmark creators further hypothesize on why natural yes/no questions (collected by the benchmark) require inference (p.2928):

- Factoid questions are rare, perhaps because people tend to phrase such questions as short-answer questions.
- "both the passages and questions rarely include negation. As a result, detecting a "no" answer typically requires understanding that a positive assertion in the text excludes, or makes unlikely, a positive assertion in the question. This requires reasoning that goes beyond paraphrasing"

Q8 - SUPPORT: What evidence do the benchmark creators offer to support *content validity* of the exanokes? In other words, we question whether the data captures capabilities of interest. Content validity is often based on analysis by external experts or benchmark users.

Nothing mentioned.

The Adaptation Module

When evaluating humans, the benchmark might instruct them to perform a task by providing instructions, training exercises, demonstrations, etc. When evaluating models/systems, there are also myriad methods that i) modify the models/systems (e.g., fine-tuning), or ii) format or add onto the input (e.g., adding examples in few-shot prompting). These adaptation methods should be chosen carefully so as to not confound evaluation results.

Q9 - DESCRIBE: Given an input, how are the objects of evaluation adapted or instructed to provide the output?

For models, the benchmark did not prescribe an adaptation method.

Note that the authors experimented with several model training strategies: directly training on BoolQ training set, finetuning, unsupervised pre-training with language modeling objective, transfer learning using various datasets, etc. These could be interpreted as adaptation methods. However, since one of the experiments in the BoolQ paper is about which strategies result in models with better IR, we consider a “model” (i.e., object of evaluation) to include its training strategy.

E.g.:
Option A: BERT finetuned on BoolQ training data = a “model” → the option we go with.
Option B: BERT = a “model”; Finetuning on BoolQ training data = presentation method.

For human annotators, there is no information on exactly what the human annotators see as stimulus (e.g., how the question and passage are presented)

Q10 - JUSTIFY: Elaborate on the suitability of the adaptation methods for all intended objects of evaluation.

Not applicable

Q11 - SUPPORT: What validity evidence do benchmark designers offer that supports the choice of the adaptation methods?

Not applicable

The Assembly Module

Items specified by the content module are what the benchmark could use. The assembly module concerns what test items from that pool will actually be used by the benchmark for evaluation, and whether this set allows the benchmark to gather sufficient evidence.

Q12 - DESCRIBE: How many items are chosen to assemble the subset used for evaluation? What factors inform this selection?

The entirety of BoolQ test set (3.2K) is used for evaluation of models. It is not clear how data points are selected from the total set (16K) to form this test set, apart from ensuring that questions from Natural Questions (prior work) are not in the test set.

For the estimation of human performance, 110 questions are randomly chosen from the total BoolQ set (minus Natural Questions).

Q13 - JUSTIFY: How does the described assembly method ensure that the produced subset elicits sufficient evidence for all capabilities of interest?

Nothing mentioned.

Q14 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of assembly methods?

Nothing mentioned.

The Evidence Module

Evidence Extraction

In response to each presented test item, objects of evaluation produce observable behaviors (referred to as “responses”) which are captured by the benchmark. From these responses, the benchmark extracts evidence about capabilities of interest that said test item targets (referred to as “salient evidence”).

Q15 - DESCRIBE: For each test item…

- **What responses are captured and used for evidence extraction?** When evaluating humans, many types of responses can be captured: selection in multiple-choice questions, long-form answers, response time, etc. Similarly, the benchmark can use the generated text (decoded in a certain way), token probabilities, running time, etc.

  Produced class label: “Yes” or “No”

  For models, the class label is obtained by the final softmax layer prediction.
  For humans, it is the annotated label that is taken as results.

- **How is evidence extracted and represented?**

  Whether the predicted/annotated label is correct (i.e. matches with the reference label):
  For models, using answers by annotators as reference.
  For humans, using answers by authors as reference.

Q16 - JUSTIFY: How does the extracted evidence capture the capabilities of interest?

Both annotators’ answers and authors’ answers are presented as good-quality labels that can be considered as ground-truths. No justification explicitly given for the metric (i.e. exact match) → Implicit? A correct label is evidence that the evaluatee is able to answer naturally yes/no questions.

Q17 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence extraction method?

Annotation quality for human annotators estimated at 90%, which supports the use of annotated answers as reference. No other evidence mentioned.

Evidence Accumulation

Q18 - DESCRIBE: How is the evidence accumulated to draw insights about the objects of evaluation in terms of capabilities of interest?

Accuracy

Q19 - JUSTIFY: How does the method of accumulating evidence capture capabilities of interest?

Nothing mentioned.

Q20 - SUPPORT: What validity evidence do the benchmark creators offer to support the choice of evidence accumulation method?

Nothing mentioned.