Gmail

Christian Richter <chrisrichter145@gmail.com>

Human Evaluation Datasheet v1.0

1 Nachricht

Google Formulare <forms-receipts-noreply@google.com> 12. August 2021 um 13:53
An: chrisrichter145@gmail.com

Google Forms

Danke, dass Sie das Formular Human Evaluation Datasheet v1.0 ausgefüllt haben

Hier sehen Sie die empfangenen Antworten.

Antwort bearbeiten

Human Evaluation Datasheet v1.0

This is a template for describing a single human evaluation experiment; it uses check-box and multiple-choice questions where possible, for comparability across experiments.
A single human evaluation experiment in this context is one that evaluates a single set of directly comparable systems in a single experimental design, but may assess multiple quality criteria.
We refer to the single human evaluation experiment that the template is being completed for simply as 'the evaluation experiment' or 'the experiment' in questions below.
The Human Evaluation Datasheet (this template) is divided into 5 sections: (1) Paper and Resources, (2) Evaluated System, (3) Quality Criteria, (4) Experimental Design, and (5) Ethics.
This sheet was developed for first use in the ReproGen Shared Task on Reproducibility of Human Evaluations in NLG, but it is applicable to all human evaluations across NLP. Some questions in version 1.0 are geared towards systems that produce language as output, but have 'Other' options for entering information for other types of systems. For version 2.0 we will increase generality.
Note that the sheet can be edited after submission: once the first section has been completed it's possible to save partial progress and continue later. The last question below is for confirming that the current submitted version of a sheet is the final one.
There are 5 or more form sections to complete: if the experiment has one quality criterion, then there are 5 sections, if it has 2 then there are 6, and so on.
Some explanation of questions and guidance for answering them is provided in this form. However, more complete explanations and guidance can be found in Shimorina & Belz, 2021 (ref). The intention is that the form is completed while referring to the paper.

Acknowledgments:
Questions relating to quality criteria in Section 4, and some of the experimental design questions in Section 3 are based on Belz et al., 2020 (https://www.aclweb.org/anthology/2020.inlg-1.24.pdf). Questions 2.1-2.5 relating to system, and 4.3.1-4.3.8 relating to response elicitation, are based on Howcroft et al., 2020 (https://www.aclweb.org/anthology/2020.inlg-1.23.pdf).
The sheet was also informed by van der Lee et al., 2019 (https://www.aclweb.org/anthology/W19-8643/); van der Lee et al., 2021 (https://www.sciencedirect.com/science/article/pii/S088523082030084X); and Gehrmann et al., 2021 (https://gem-benchmark.com/data_cards/guide).
More generally, the inspiration for creating a 'sheet' for describing human evalution experiments of course came from Gebru et al., 2018 (https://arxiv.org/pdf/1803.09010.pdf); Bender & Friedman, 2018 (https://www.mitpressjournals.org/doi/pdfplus/10.1162/tacl_a_00041); and Mitchell et al., 2019 (https://dl.acm.org/doi/abs/10.1145/3287560.3287596).

E-Mail-Adresse *

chrisrichter145@gmail.com

1. PAPER AND RESOURCES

This section records information about contact person, paper reporting the experiment, and resources used in it, unless the sheet is being completed for preregistration.

1.1 Link to paper reporting the evaluation experiment. If the paper reports more than one experiment, state which experiment you're completing this sheet for. Or, if applicable, enter 'for preregistration.' *

https://aclanthology.org/W18-6532/

1.2 Link to website providing resources used in the evaluation experiment (e.g. system outputs, evaluation tools, etc.). If there isn't one, enter 'N/A'. *

https://github.com/der-Richter/ReproGen_HumanEvaluationReproduction

1.3 Name, affiliation and email address of person completing this sheet, and of contact author if different. *

Christian Richter, chrisrichter145@gmail.com, Technical University of Darmstadt (Student) &
Yanran Chen, chenyr1996@hotmail.com, Technical University of Darmstadt (Student)

2. EVALUATED SYSTEM(S)

This section records information about the system(s) (or human stand-ins) whose outputs are being evaluated in the evaluation experiment that this sheet is being completed for.

2.1 What type of input do the evaluated system(s) take? Select all that apply. If none match, select `Other' and describe.

- [x] raw/structured data
- [ ] deep linguistic representation (DLR)
- [ ] shallow linguistic representation (SLR)
- [ ] text: subsentential units of text
- [ ] text: sentence
- [ ] text: multiple sentences
- [x] text: document
- [ ] text: dialogue
- [ ] text: other
- [ ] speech
- [ ] visual
- [ ] multi-modal
- [ ] control feature
- [ ] no input (human generation)
- [ ] Sonstiges:

2.2 What type of output do the evaluated system(s) generate? Select all that apply. If none match, select `Other' and describe.

- [ ] raw/structured data
- [ ] deep linguistic representation (DLR)
- [ ] shallow linguistic representation (SLR)
- [ ] text: subsentential unit of text
- [ ] text: sentence
- [x] text: multiple sentences
- [ ] text: document
- [ ] text: dialogue
- [ ] text: other
- [ ] speech
- [ ] visual
- [ ] multi-modal
- [ ] human-generated ‘outputs’
- [ ] Sonstiges: .....

2.3 How would you describe the task that the evaluated system(s) perform in mapping the inputs in Q2.1 to the outputs in Q2.2? Occasionally, more than one of the options below may apply. If none match, select `Other' and describe.

- [ ] content selection/determination
- [ ] content ordering/structuring
- [ ] aggregation
- [ ] referring expression generation

- [ ] lexicalisation
- [ ] deep generation
- [ ] surface realisation (SLR to text)
- [ ] feature-controlled text generation
- [ ] data-to-text generation
- [ ] dialogue turn generation
- [ ] question generation
- [ ] question answering
- [ ] paraphrasing / lossless simplification
- [ ] compression / lossy simplification
- [ ] machine translation
- [x] summarisation (text-to-text)
- [ ] end-to-end text generation
- [ ] image/video description
- [ ] post-editing/correction
- [ ] Sonstiges:

2.4 Input language(s), or 'N/A'.

English

2.5 Output language(s), or 'N/A'.

English

3. OUTPUT SAMPLE, EVALUATORS, EXPERIMENTAL DESIGN

This section records information about the sample of outputs, the evaluators and the experimental design used in the experiment.

3.1 SAMPLE OF SYSTEM OUTPUTS (OR HUMAN-AUTHORED STAND-INS) EVALUATED

This subsection records information about the system outputs evaluated in the human evaluation experiment.

3.1.1 How many system outputs (or other evaluation items) are evaluated per system in the evaluation experiment? Answer should be an integer.

30

3.1.2 How are system outputs (or other evaluation items) selected for inclusion in the evaluation experiment? If none match, select `Other' and describe.

- (●) by an automatic random process from a larger set
- ( ) by an automatic random process but using stratified sampling over given properties
- ( ) by manual, arbitrary selection
- ( ) by manual selection aimed at achieving balance or variety relative to given properties
- ( ) Sonstiges:

3.1.3 What is the statistical power of the sample size? See e.g. Card et al., 2020 (https://www.aclweb.org/anthology/2020.emnlp-main.745/).

3.2 EVALUATORS

This subsection records information about the evaluators participating in the evaluation experiment.

3.2.1 How many evaluators are there in this experiment? Answer should be an integer.

19

3.2.2 What kind of evaluators are in this experiment? Select all that apply. If none match, select `Other' and describe. In all cases, provide details in the text box under `Other'.

- [ ] experts
- [x] non-experts
- [ ] paid (including non-monetary compensation such as course credits)
- [x] not paid
- [x] previously known to authors
- [x] not previously known to authors
- [ ] evaluators include one or more of the authors
- [x] evaluators do not include any of the authors
- [x] Sonstiges: Not previously known to all authors

3.2.3 How are evaluators recruited?

By selecting from the circle of acquaintances, reduced to english-speakers.

3.2.4 What training and/or practice are evaluators given before starting on the evaluation itself?

One Example and a quick guide

3.2.5 What other characteristics do the evaluators have, known either because these were qualifying criteria, or from information gathered as part of the evaluation?

3.3 EXPERIMENTAL DESIGN

This subsection records information about the experimental design of the human evaluation experiment.

3.3.1 Has the experimental design been preregistered? If yes, on which registry?

No

3.3.2 How are responses collected? E.g. paper forms, online survey tool, etc.

Google Form - Online Survey

3.3.3 What quality assurance methods are used? Select all that apply. If none match, select `Other' and describe. In all cases, provide details in the text box under `Other'.

- [ ] evaluators are required to be native speakers of the language they evaluate
- [ ] automatic quality checking methods are used during/post evaluation
- [x] manual quality checking methods are used during/post evaluation
- [ ] evaluators are excluded if they fail quality checks (often or badly enough)

- [x] some evaluations are excluded because of failed quality checks
- [ ] none of the above
- [x] Sonstiges: One evaluator selected only two values randomly and got replaced then.

3.3.4 What do evaluators see when carrying out evaluations? Link to screenshot(s) and/or describe the evaluation interface(s).

https://forms.gle/QG7QA7U9XHcDBVtv5

3.3.5 How free are evaluators regarding when and how quickly to carry out evaluations? Select all that apply. In all cases, provide details in the text box under `Other'.

- [ ] evaluators have to complete each individual assessment within a set time
- [x] evaluators have to complete the whole evaluation in one sitting
- [ ] neither of the above
- [ ] Sonstiges:

3.3.6 Are evaluators told they can ask questions about the evaluation and/or provide feedback? Select all that apply. In all cases, provide details in the text box under `Other'.

- [ ] evaluators are told they can ask any questions during/after receiving initial training/instructions, and before the start of the evaluation
- [ ] evaluators are told they can ask any questions during the evaluation
- [ ] evaluators are asked for feedback and/or comments after the evaluation, e.g. via an exit questionnaire or a comment box
- [x] None of the above
- [ ] Sonstiges:

3.3.7 What are the experimental conditions in which evaluators carry out the evaluations? In all cases, provide details in the text box under `Other'.

- ○ evaluation carried out by evaluators in a place of their own choosing, e.g. online, using a paper form, etc.
- ○ evaluation carried out in a lab, and conditions are the same for each evaluator
- ○ evaluation carried out in a lab, and conditions vary for different evaluators
- ○ evaluation carried out in a real-life situation, and conditions are the same for each evaluator
- ◉ evaluation carried out in a real-life situation, and conditions vary for different evaluators
- ○ evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions are the same for each evaluator
- ○ evaluation carried out outside of the lab, in a situation designed to resemble a real-life situation, and conditions vary for different evaluators
- ○ Sonstiges: .....

3.3.8 Unless the evaluation is carried out at a place of the evaluators' own choosing, briefly describe the (range of different) conditions in which evaluators carry out the evaluations.

.....

4. QUALITY CRITERION 1 - DEFINITION AND OPERATIONALISATION

This section records information about the first quality criterion assessed in the experiment. At the end of this section, a link is provided to a duplicate section for further quality criteria assessed as part of the same experiment.

4.1 QUALITY CRITERION PROPERTIES

This subsection records information about (i) what type of quality is being assessed; (ii) what aspect of the system output is being assessed; and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference.

4.1.1 What type of quality is assessed by the quality criterion? (For explanation of terms see next field below.)

- (●) Correctness
- ( ) Goodness
- ( ) Features

Explanation

1. Correctness: For correctness criteria it is possible to state, generally for all outputs, the conditions under which outputs are maximally correct (hence of maximal quality). E.g. for Grammaticality, outputs are (maximally) correct if they contain no grammatical errors; for Semantic Completeness, outputs are correct if they express all the content in the input.
2. Goodness: For goodness criteria, in contrast to correctness criteria, there is no single, general mechanism for deciding when outputs are maximally good, only for deciding for two outputs which is better and which is worse. E.g. for Fluency, even if outputs contain no disfluencies, there may be other ways in which any given output could be more fluent.
3. Features: For criteria X in this class, outputs are not generally better if they are more X. Depending on evaluation context, more X maybe better or less X may be better. E.g. outputs can be more specific or less specific, but it's not the case that outputs are, in the general case, better when they are more specific.

4.1.2 Which aspect of system outputs is assessed by the quality criterion? (For explanation of terms see next field below.)

- ( ) Form of output
- ( ) Content of output
- (●) Both form and content of output

Explanation

1. Form of output: Evaluations of this type aim to assess the form of outputs alone, e.g. Grammaticality is only about the form, a sentence can be grammatical yet be wrong or nonsensical in terms of content.
2. Content of output: Evaluations of this type aim to assess the content/meaning of the output alone, e.g. Meaning Preservation only assesses output content; two sentences can be considered to have the same meaning, but differ in form.
3. Both form and content of output: Here, evaluations assess outputs as a whole, not distinguishing form from content. E.g. Coherence is a property of outputs as a whole, either form or meaning can detract from it.

4.1.3 Is each output assessed for quality in its own right, or with reference to a

system-internal or external frame of reference? (For explanation of terms see next field below.)

- ○ Quality of output in its own right
- ◉ Quality of output relative to the input
- ○ Quality of output relative to a system-external frame of reference

Explanation

1. Quality of output in its own right: output quality is assessed without referring to anything other than the output itself, i.e. no system-internal or external frame of reference. E.g Poeticness is assessed by considering (just) the output and how poetic it is.
2. Quality of output relative to the input: the quality of an output is assessed relative to the input. E.g. Answerability is the degree to which the output question can be answered from information in the input.
3. Quality of output relative to a system-external frame of reference: output quality is assessed with reference to system-external information, such as a knowledge base, a person's individual writing style, or the performance of an embedding system. E.g. Factual Accuracy assesses outputs relative to a source of real-world knowledge.

4.2 EVALUATION MODE PROPERTIES

This subsection records information about evaluation modes which are orthogonal to quality criteria, i.e. any given quality criterion can in principle be combined with any of the modes (although some combinations are more common than others).

4.2.1 Does an individual assessment involve an objective or a subjective judgment? (For explanation of terms see next field below.)

- ○ Objective
- ◉ Subjective

Explanation

Examples of objective assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. Repeated assessments of the same output with an objective-mode evaluation method always yield the same score/result.
Subjective assessments involve ratings, opinions and preferences by evaluators. Some criteria lend themselves more readily to subjective assessments, e.g. Friendliness of a conversational agent, but an objective measure e.g. based on lexical markers is also conceivable.

4.2.2 Are outputs assessed in absolute or relative terms? (For explanation of terms

see next field below.)

- (●) Absolute
- ( ) Relative

Explanation

In absolute assessment, evaluators are shown outputs from a single system during evaluation.
In relative assessment, evaluators are shown outputs from multiple systems at the same time, typically ranking or preference-judging them.

4.2.3 Is the evaluation intrinsic or extrinsic? (For explanation of terms see next field below.)

- (●) Intrinsic
- ( ) Extrinsic

Explanation

Extrinsic evaluations assess quality of outputs in terms of their effect on something external to the system, e.g. performance of an embedding system or of a user at a task. Intrinsic evaluations do not.

4.3 RESPONSE ELICITATION

This subsection records information about how responses are elicited for the quality criterion this section is being completed for.

4.3.1 What do you call the quality criterion in explanations/interfaces to evaluators? Enter `N/A' if criterion not named.

Information Coverage, Non-Redundancy, Semantic Adequacy, Grammatical Correctness

4.3.2 What definition do you give for the quality criterion in explanations/interfaces to evaluators? Enter `N/A' if no definition given.

N/A

4.3.3 Size of scale or other rating instrument (i.e. how many different possible values there are). Answer should be an integer or 'continuous' (if it's not possible to state how many possible responses there are). Enter 'N/A' if there is no rating instrument.

5

4.3.4 List or range of possible values of the scale or other rating instrument. Enter 'N/A', if there is no rating instrument.

worst to best: 1, 2, 3, 4, 5

4.3.5 How is the scale or other rating instrument presented to evaluators? Use the textbox under `Other' to provide details and link to a screenshot.

- ○ multiple-choice options
- ◉ check-boxes
- ○ slider
- ○ N/A (there is no rating instrument)
- ○ Sonstiges:

4.3.6 If there is no rating instrument, describe briefly what task the evaluators perform (e.g. ranking multiple outputs, finding information, playing a game, etc.), and what information is recorded. Enter 'N/A' if there is a rating instrument.

N/A

4.3.7 What is the verbatim question, prompt or instruction given to evaluators (visible to them during each individual assessment)?

How do you judge the <CRITERIA> of the company summary?

4.3.8 Form of response elicitation. If none match, select `Other' and describe.

- ○ (dis)agreement with quality statement
- ◉ direct quality estimation
- ○ relative quality estimation (including ranking)
- ○ counting occurrences in text
- ○ qualitative feedback (e.g. via comments entered in a text box)
- ○ evaluation through post-editing/annotation
- ○ output classification or labelling
- ○ user-text interaction measurements
- ○ task performance measurements
- ○ user-system interaction measurements
- ○ Sonstiges:

4.3.9 How are raw responses from participants aggregated or otherwise processed to obtain reported scores for this quality criterion? State if no scores reported.

Formation of average values per system and additional calculation of correlations by means of

spearman's

4.3.10 Method(s) for determining effect size and significance of findings for this quality criterion.

spearman's correlation

4.3.11 Has the inter-annotator and intra-annotator agreement between evaluators for this quality criterion been measured? If yes, what method was used, and what are the agreement scores?

No

Is there another quality criterion in the evaluation experiment that you haven't completed this section for yet? *

- ○ Yes
- ◉ No

5. Ethics

This section records information of a broadly ethical nature.

5.1 Has the evaluation experiment this sheet is being completed for, or the larger study it is part of, been approved by a research ethics committee? If yes, which research ethics committee?

no

5.2 Do any of the system outputs (or human-authored stand-ins) evaluated, or do

any of the responses collected, in the experiment contain personal data (as defined in GDPR Art. 4, §1: https://gdpr.eu/article-4-definitions/)? If yes, describe data and state how addressed.

no

5.3 Do any of the system outputs (or human-authored stand-ins) evaluated, or do any of the responses collected, in the experiment contain special category information (as defined in GDPR Art. 9, §1: https://gdpr.eu/article-9-processing-special-categories-of-personal-data-prohibited/)? If yes, describe data and state how addressed.

no

5.4 Have any impact assessments been carried out for the evaluation experiment, or the larger study it is part of, and/or any data collected/evaluated in connection with it? If yes, summarise approach and outcomes.

Common sense, no impact seen

Closing

You have reached the end of the Human Evaluation Sheet. Is the information as entered complete and final? *

- (●) Yes
- ( ) No, I will need to come back and add/change information