Evaluation of Response Generation Models: Shouldn't It Be Shareable and Replicable?

S. **Mahed** Mousavi, Gabriel Roccabruna, Michela Lorandi, Simone Caldarella, Giuseppe Riccardi

Signals and Interactive Systems Lab, University of Trento, Italy

Official repo

Introduction

Automatic metrics are **inadequate** & **correlate poorly** with human judgement.

Most papers publish either exclusively automatic metrics or **incomplete, not-comparable & not interpretable Human Evaluations** (HE).

Could we facilitate a convergence towards a de-facto standard or at least fully disclosed & publicly-shared HE Protocol?

Approach

We propose to **standardize the HE** task of response generation models by **publicly sharing a detailed protocol**:

a. *domain-agnostic*,
b. *language-independent*,
c. *extendable*

Public release the protocol & its materials including GUI, guidelines, & scripts used.

We invite the community to **utilize this protocol *AND*** to **improve & extend** it into **referable** & **version-able standards**.

Step 1: Task Design

Evaluation Aspect

1. Evaluation Granularity
2. Quality Dimensions
3. Questions
4. Decisions
5. Explanations

Annotation Aspect

Main causes of error:
a. *Cognitive Workload*
b. Mental Model *Gaps*

1. Guidelines & Examples
2. User Interface
3. Internal Pilots

Step 2: Annotator Recruiting

Relevant recruiting aspects:

1. Sampling
2. Qualification
3. Compensation

Step 3: Task Execution

- Continuous control
- Inter-Annotator Agreement level (IAA)
- Real-time feedback

Step 4: Annotation Reporting

Checklist of aspects and elements to facilitate **replicability & reproducibility**

Evaluation
- Granularity
- Dimensions
- Questions
...

Annotation
- #Annotators
- #Votes
- IAA
...

Recruitment
- Sampling
- Qualification
- #Workers

HE Protocol Validation

Models: GPT-2 & T5,
Task: *personal & grounded* response generation.

Annotators: 35
Dialogues: 42 (1 dialogue : 7 annotators)

Observations:
a. **huge gap** to reach the quality of ground truth.
b. **High Subjectivity** in *Contextualization* & *Listening*.

Effect of Grounding
a. reduces ***Genericness*** percentage.
b. increases ***Hallucination*** percentage.
c. does not affect ***Incoherency***.

Models	Inter Annotator Agreement Level measured by Fleiss' κ				
	Appropriateness	*Contextualization*	*Correctness*	*Listening*	IAA per Model
GePpeTto	0.27	0.14	0.64	0.15	0.32±0.10
+Knowledge	0.42	0.22	0.36	0.27	0.36±0.11
iT5-Base	0.24	0.19	0.06	0.18	0.27±0.04
+Knowledge	0.18	0.03	0.30	0.21	0.19±0.06
IAA per Dimension	0.30 ±0.10 Fair	0.15±0.05 Poor	0.41±0.20 Moderate	0.23±0.07 Fair	-

Models	Human Evaluation			
	Appropriateness	Contextualization	Correctness	Listening
Ground Truth	100.0%	97.62%	97.62%	97.62%
GePpeTto	66.67%	69.05%	83.33%	64.29%
+Knowledge	59.52%	57.14%	83.33%	57.14%
iT5-Base	66.67%	73.81%	100.0%	66.67%
+Knowledge	80.95%	80.95%	85.71%	76.19%