QCET: An Interactive Taxonomy of Quality Criteria for Comparable and Repeatable Evaluation of NLP Systems

Anya Belz, Simon Mille, Craig Thomson and Rudali Huidrom

DCU Natural Language Generation Research Group ADAPT, Dublin City University, Dublin, Ireland anya.belz@dcu.ie

Abstract

Four years on from two papers (Belz et al., 2020; Howcroft et al., 2020) that first called out lack of standardisation and comparability in quality criteria assessed in NLP system evaluations, researchers still use widely differing quality criteria names and definitions, meaning that it continues to be unclear when the same aspect of quality is being assessed in two evaluations. While normalised quality criteria were proposed at the time, the list was unwieldy and using it came with a steep learning curve. In this demo paper, our aim is to address these issues with an interactive taxonomy tool that enables quick perusal and selection of the standardised quality criteria, and provides decision support and examples of use at each node.

1 Introduction

Belz et al. (2020) and Howcroft et al. (2020) described a situation in NLP evaluation where over 200 different quality criteria (QC) names were in use, definitions were patchy, and it was impossible to tell if the same aspect of quality was being evaluated in different evaluations, resulting in low comparability and repeatability. Consider the following examples which share the same QC name, **Fluency**, but use very different definitions:

- 1. Yu et al. (2020): "judging the question fluency."
- 2. Van de Cruys (2020): "grammatical and syntactically well-formed."
- 3. Pan et al. (2020): "follows the grammar and accords with the correct logic."

The authors mapped the 200+ different QCs to a set of 71 standardised QC names and definitions. This implies that two thirds of the original set did not reflect actual differences between aspects of quality assessed, merely a lack of standardisation.

While highly cited, the work has found little practical application in that researchers continue to create new names and/or definitions for the aspects of quality they assess. We diagnose two reasons for this: (i) it is not straightforward to separate *what* is being assessed from *how* it is being assessed; and (ii) the 71 QCs were difficult to assimilate and use.

2 Disentangling the *What* from the *How*

What is being assessed refers to the specific aspect of quality (the QC) that an evaluation aims to assess. *How* refers to the way in which the QC is mapped to a specific measure that can be implemented in an evaluation method. The distinction matters, because there are many different ways in which the same quality criterion can be assessed (Belz et al., 2020). The different elements relate as follows (*evaluation mode*: intrinsic vs. extrinsic, subjective vs. objective and absolute vs. relative):

Quality criterion + evaluation mode = evaluation measure; Evaluation measure + experimental design = evaluation method.

3 The Underlying QC Taxonomy

The root node is the single most general QC, Overall Quality. The next level down has the following three QC subclasses (Belz et al., 2020):

- 1. **Correctness**: The conditions under which outputs are maximally correct (hence of maximal quality) can be stated. E.g. for *Grammaticality*, outputs are (maximally) correct if they contain no grammatical errors.
- 2. **Goodness**: It cannot be stated when outputs are maximally good, only which of two is better/worse. E.g. for *Fluency*, even if an output contains no disfluencies, there may be other ways to improve its fluency.
- 3. Features: For a feature-type QC X, outputs are not generally better if they are more (or less) X. Depending on context, more X may be better or worse. E.g. *Funny* and *Entertaining* might be good in a narrative generator, but neither are appropriate in a nursing report generator.

At the next level, these three nodes split into subclasses that reflect whether outputs are assessed in their own right, relative to the input, or relative to an external frame of reference (e.g. comparison to a gold standard).

At a subsequent level, classes further split into the following three subclasses capturing which aspect of an output is being assessed:

- 1. **Form**: The form of outputs alone is assessed, e.g. *Grammaticality* a sentence can be grammatical yet wrong or nonsensical in terms of content.
- 2. **Content**: The content/meaning of outputs alone is assessed, e.g. *Meaning Preservation* two sentences can have the same meaning, but differ in form.
- 3. **Both form and content**: Outputs are assessed as a whole. E.g. *Coherence* is a property of outputs as a whole, either form or meaning can detract from it.

The rest of the taxonomy consists of 70+ leaf and internal nodes; the following is an example of a complete branch from root to leaf node:

Overall quality of outputs \rightarrow Correctness of outputs \rightarrow Correctness of outputs in their own right \rightarrow Correctness of outputs in their own right (Form) \rightarrow Grammaticality.

4 The QCET Tool

The Quality Criteria for Evaluations Taxonomy (QCET) tool has three core functionalities: (i) tree navigation via *show child nodes* (+) and *collapse subtree* (-); (ii) tree pruning to one of the subclasses in Section 3; (iii) viewing node details.

The starting interface (Appendix A) has instructions corresponding to Section 3 and the use cases below, followed by the three tree pruning pulldowns, and the taxonomy viewer, initially showing just the root node (*Quality of Outputs*).

The taxonomy tree can then be navigated by progressively showing child nodes and collapsing subtrees. Alternatively, to provide less comprehensive views, the taxonomy tree can be pruned in the following three ways:

1. **Prune by Level 1 QC classes:** show only one of the three Level 1 subtrees, corresponding to showing *Correctness* QCs, *Goodness* QCs, or *Feature-type* QCs only. Options are selected via the following pull-down:

✓ Show all Level 1 QC Classes
[C] Show Correctness QCs only
[G] Show Goodness QCs only
[F] Show Feature-type QCs only

- 2. Prune by Level 2 QC classes: show all and only subtrees rooted at *one* of the following: a QC evaluating outputs *In their Own Right*, a QC evaluating outputs *Relative to the Inputs*, or QC evaluating outputs *Relative to an External Frame of Reference*.
- 3. **Prune by Form/Content/Both QC classes**: show all and only subtrees rooted at *one* of the following: a *Form* QC, a *Content* QC, or a *Form and Content* QC.

The three pruning dimensions can be combined, so that e.g. selecting *Correctness QCs only* and *Form QCs only* at the same time will show just subtrees that are in both these classes.

Nodes. Each *QC node* has the following elements: (i) node ID, (ii) QC name; (iii) QC definition; (iv) suggested questions to put to the evaluators in a corresponding human evaluation of this



Figure 1: Example node with details shown.

criterion (a) in absolute and (b) relative evaluation mode; and (v) additional notes and information.

The node ID traces the path from the root to the node, e.g. QCO-f: Quality \rightarrow Correctness \rightarrow In its own right \rightarrow Form. Each node has a (+) button or a (-) button, standing for *show child nodes* and *collapse subtree*, respectively. Nodes by default show just the QC name and definition; this view can be expanded in situ, by clicking on *show details*, or by clicking on the magnifying class icon which opens the full view of the node details in a new tab.

See Figure 1 for an example of a non-expanded node viewed with details shown. Appendix B shows the expanded view of a node.

Use cases. We envisage two main uses: (i) Mapping previous evaluations to standardised QCs for comparability, and (ii) choosing a QC name and definition for new evaluations. In both cases, the default use mode is to peruse the taxonomy from the root node down until the right node is found (more details in Appendix C). For users more familiar with the taxonomy, the tree-pruning pull-downs offer a convenient way to reduce the search space.

Extensibility. The QCET tool is intended as an extensible resource, to which new leaf nodes can be added as needed. The tool includes an interface for proposing new nodes to be inserted at a specific parent node in the publicly shared version which the QCET team will review and, by default, add.

5 Conclusion

We have presented QCET,¹ a tool designed to make it easier to (i) identify which standardised QC is being assessed in an existing evaluation, and (ii) to select standard names and definitions to use in new evaluations, in order to increase comparability and repeatability of evaluation experiments overall.

¹https://github.com/DCU-NLG/qcet

References

- Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In Proceedings of the 13th International Conference on Natural Language Generation, pages 183-194, Dublin, Ireland. Association for Computational Linguistics.
- David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid Hasan, Saad Mahamood, Simon Mille, Sashank Santhanam, Emiel van Miltenburg, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Proceedings of the 13th International Natural Language Generation Conference.
- Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1463-1475, Online. Association for Computational Linguistics.
- Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2471–2480.
- Qian Yu, Lidong Bing, Qiong Zhang, Wai Lam, and Luo Si. 2020. Review-based question generation with adaptive instance transfer and augmentation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 280–290, Online. Association for Computational Linguistics.

Α **QCET Starting Interface**

Figure 2 shows the QCET starting interface. The user can view the instructions under the 5 headers at the top, before pruning the Taxonomy Tree with the dropdown menus and then expanding/collapsing nodes and their details on the tree itself.

B **Single Node View**

Figure 3 shows the details for node QCO-f; Correctness of outputs in their own right (form). In addition to showing the definition, the Node Details show suggested questions for different evaluation modes. Also shown is the parent node QCO and the child nodes QCO-f-1 and 0C0-f-2.

Figure 4 shows details for the leaf node QCO-f-1; Grammaticality, the first child node of QCO-f.

Taxonomy of Quality Criteria For Evaluations (QCET) Tool

Instructions

Click on the headers below to show the instructions

The Underlying QC Taxonomy	\sim
The QCET Tool	\sim
Nodes	\sim
Use cases	\sim
Extensibility	\sim

Taxonomy Tree

Prune tree by level 1 QC classes (Correctness, Goodness, Features) Show all Level 1 QC Classes ~

Prune tree by level 2 QC classes (In its own Right, Relative to Input, Relative to External Frame of Reference)

Show all Level 2 QC Classes	•	
Prune tree by form vs. content QC classes (Form, Content, Both)		
Show all Form/Content/Both QC Classes	~	

$\left(\right)$	Q : Quality of outputs
	The overall quality of an output. 📳
	show details

Figure 2: Example node in single view; Correctness of outputs in their own right (form).

Showing Node OCO-f

	<u>QCO : Correctness of outputs in their own right</u>
lo	de Details
	QCO-f : Correctness of outputs in their own right (form)
	Definition:
	The degree to which an output is correct, considering only the output,
	and assessed on its form only.
	Suggested question to evaluators in subjective, intrinsic, absolute
	mode:
	To what degree is this output correct, looking at its form only and
	ignoring its content/meaning?
	Suggested question to evaluators in subjective, intrinsic, relative mode:
	Which of these outputs is more correct, looking at its form only and
	ignoring its content/meaning?
	Additional notes and information:
	None.

hildre	n	
$\left(\right)$	<u>QCO-f-1 : Grammaticality</u>	
$\left(\right)$	<u>QCO-f-2 : Spelling accuracy</u>	

Figure 3: Example node in single view; Correctness of outputs in their own right (form).

Showing Node QCO-f-1





Figure 4: Example node in single view; Grammaticality (child node of QCO-f).

C Example Uses

Identifying standardised QCs in existing experiments. The first step is to locate all resources shared about a given experiment, then to identify (i) QC name, (ii) QC definition and (iii) the question and/or instructions put to evaluators. In many cases, (i)–(iii) are not completely aligned in which case (iii) takes priority as expressing what was actually evaluated.

A complicating factor is often that in the effort of explaining one QC, developers often introduce terms associated with other QCs, e.g. in the second Fluency definition at the start of the paper, Fluency is explained (only) in terms of grammaticality which introduces another QC. The third definition introduces two other QCs. We would argue that in the former case, one QC is being evaluated (Grammaticality), and in the latter case two (Grammaticality and Logicality). Neither assesses Fluency.

To arrive at these conclusions, armed with the information above, the taxonomy is perused via the QCET tool top down until the correct node(s) is/are reached (corresponding to the specific individual quality criteria above).

Designing new evaluation experiments. Once an initial idea has been formed about what aspect to assess in the new evaluation, the taxonomy is perused top down until the correct node(s) is/are reached. What is not supplied by the taxonomy is what explanations and/or instructions to issue to evaluators. Our recommendation would be to avoid paraphrasing the QC name and to rely more on examples for each response value. E.g. if a 5point rating scale is used for Fluency, then provide a set of examples for each point on the scale. We are planning to add standardised questions for each node in the future.