Disentangling the Properties of Human Evaluation Methods: A Classification System to Support Comparability, Meta-Evaluation and Reproducibility Testing

Anya Belz
University of Brighton, UK
a.s.belz@brighton.ac.uk

Simon Mille
UPF, Barcelona, Spain
simon.mille@upf.edu

David M. Howcroft
Heriot-Watt University, UK
d.howcroft@hw.ac.uk

Abstract

Current standards for designing and reporting human evaluations in NLP mean it is generally unclear which evaluations are comparable and can be expected to yield similar results when applied to the same system outputs. This has serious implications for reproducibility testing and meta-evaluation, in particular given that human evaluation is considered the gold standard against which the trustworthiness of automatic metrics is gauged. Using examples from NLG, we propose a classification system for evaluations based on disentangling (i) *what* is being evaluated (which aspect of quality), and (ii) *how* it is evaluated in specific (a) evaluation modes and (b) experimental designs. We show that this approach provides a basis for determining comparability, hence for comparison of evaluations across papers, meta-evaluation experiments, reproducibility testing.

1 Introduction

Human evaluations play a central role in Natural Language Generation (NLG), a field which has always been wary of automatic evaluation metrics and their limitations (Reiter and Belz, 2009; Novikova et al., 2017; Reiter, 2018). NLG has trusted human evaluations perhaps more than any other NLP subfield, and has always gauged the trustworthiness of automatic evaluation metrics in terms of how well, and how consistently, they correlate with human evaluation scores (Over et al., 2007; Gatt and Belz, 2008; Bojar et al., 2016; Shimorina et al., 2018; Ma et al., 2019; Mille et al., 2019; Dušek et al., 2020). If they do not, even in isolated cases, the reliability of the metric is seen as doubtful, regardless of the quality of the human evaluation, or whether the metric and human evaluation involved aimed to assess the same thing.

More generalised conclusions are sometimes drawn, for example that BLEU scores do not correlate well with human judgements of specific quality criteria[1] such as 'Fluency,' 'Naturalness,' 'Readability' or 'Overall Quality'[2] in the general case (Novikova et al., 2017; May and Priyadarshi, 2017; Reiter, 2018; Shimorina et al., 2018; Dušek et al., 2020; Sellam et al., 2020; Mathur et al., 2020). However, such comments make the assumption that, and only really make sense if, multiple evaluations of, say, 'Fluency' do in fact assess the same aspect of quality in the output texts. We argue that we do not currently have a way of establishing whether any two evaluations, metric or human, do or do not assess the same thing.

In fact, we have plenty of evidence (Section 2) that in many cases, when two evaluations use the same name for a quality criterion, they do in fact assess different aspects of quality, even for seemingly straightforward criteria like 'Fluency' and 'Readability.' And conversely, evaluations that do use different terms often assess identical aspects of quality. In this situation, not only are we on shaky ground when drawing conclusions from meta-evaluations of metrics via correlations with human evaluations, but not knowing when two different evaluations should produce the same results also has clear implications for reproducibility assessments.

In this paper, we propose a classification system that disentangles the properties of evaluation methods, providing a basis for establishing comparability. We start with issues in how human evaluations are currently designed and reported in NLG (Section 2). We then discuss the difficulties of disentangling the properties of evaluation methods (Section 3), and present the proposed classification system consisting of three quality-criterion properties, three evaluation modes, and 12 experimental design properties (Section 4). Next we demonstrate how these combine to form a classification system that supports comparability (Section 5), and show how the system can be used in the context of de-

[1] Term initially used informally, defined in Section 3.
[2] Quotes to indicate no specific meaning intended.

Proceedings of The 13th International Conference on Natural Language Generation, pages 183–194,
Dublin, Ireland, 15-18 December, 2020.

signing and reporting evaluations, meta-evaluations and reproducibility testing (Section 6). We finish with some discussion and conclusions (Section 7).

Notational conventions: We use boldface for defined terms where they are being defined (e.g. **quality criterion**), italics where we want to emphasise that we are using a term in its defined meaning (e.g. *quality criterion*), and normal font otherwise; a combination of italics, boldface and capitalised initials for names of quality criteria with definitions (e.g. ***Fluency***); and italics and double quotes for verbatim definitions of quality criteria from papers (e.g. *"ease of reading"*).

2 Issues in Comparing Human Evaluations in NLG

Human evaluations in NLG currently paint a confused picture[3] with very poor standards for designing and reporting evaluations (van der Lee et al., 2019). In this section we focus on those aspects that make it hard to compare different evaluations.

2.1 Quality criterion names

Different papers use the same quality criterion name with different definitions, and the same definitions with different names. Even for less problematic criteria names such as ***Readability***,[4] substantial variation exists. Some definitions are about reading ease: *"Ease of reading"* (Forrest et al., 2018); *"a summary is readable if it is easy to read and understand"* (Di Fabbrizio et al., 2014). Others veer towards fluency: *"how fluent and readable [the text is]"* (Belz and Kow, 2010); *"readability concerns fluency of the textual data"* (Mahapatra et al., 2016). Yet others combine multiple aspects of quality: *"measures the linguistic quality of text and helps quantify the difficulty of understanding the text for a reader"* (Santhanam and Shaikh, 2019); *"[r]eadability is [...] concerned with the fluency and coherence of the texts."* (Zang and Wan, 2017).

A far messier criterion name is ***Coherence***, some definitions referring to structure (underlined text below) and theme/topic (dotted underline), some just to one of the two, and others to neither (last three examples): *"[whether] the poem [is] thematically structured"* (Van de Cruys, 2020); *"measures if a question is coherent with previous ones"* (Chai and Wan, 2020); *"measures ability of the dialogue system to produce responses consistent with the topic of conversation"* (Santhanam and Shaikh, 2019); *"measures how much the response is comprehensible and relevant to a user's request"* (Yi et al., 2019); *"refers to the meaning of the generated sentence, so that a sentence with no meaning would be rated with a 1 and a sentence with a full meaning would be rated with a 5"* (Barros et al., 2017); *"measures [a conversation's] grammaticality and fluency"* (Juraska et al., 2019); *"concerns coherence and readability"* (Murray et al., 2010).

The inverse is also common, where the same definition is used with different criterion names. E.g. Chen et al. (2020) define ***Language Naturalness*** as *"whether the generated text is grammatically correct and fluent, regardless of factual correctness"*, while Juraska et al. (2019) give essentially the same definition (see preceding paragraph) for ***Coherence***. Wubben et al. (2016) define ***Fluency*** as *"the extent to which a sentence is in proper, grammatical English"*, while Harrison and Walker (2018) use a very similar definition for ***Grammaticality***: *"adherence to rules of syntax, use of the wrong wh-word, verb tense consistency, and overall legitimacy as an English sentence."*

In some cases where criterion names are different, it is slightly more evident that criteria are in fact closely related, as with Wang et al. (2020)'s ***Faithfulness***, Cao et al. (2020)'s ***Content similarity***, and Zhou et al. (2020)'s ***Content preservation***, all of which measure the extent to which the content of an output overlaps with that of the input.

However, in many cases similarities are unguessably obscured behind criteria names, as is the case for the following names, all defined as the usefulness of the output text for completing a particular task: ***Dialogue efficiency*** (Qu and Green, 2002), ***Usefulness*** (Miliaev et al., 2003), ***Task completion*** (Varges, 2006), ***Productivity*** (Allman et al., 2012).

2.2 Other aspects of evaluations

A vanishingly small number of papers provide full details of human evaluation experiments. It is common for papers not to report how many system outputs or evaluators were used, what information was given to them, what questions asked, etc. Our survey of 468 individual human evaluations in NLG (Howcroft et al., 2020) indicates that in about 2/3 of cases reports do not provide the question/prompt evaluators were shown, over half do not define the

[3] See our survey of 20 years of human evaluations in NLG (Howcroft et al., 2020).

[4] Note that the examples in this section were chosen at random, not because they vary most widely.

quality criterion assessed, and around 1/5 do not name the quality criterion. Missing information about experimental design is particularly problematic for reproducibility testing (Section 6.3).

While some aspects of evaluations such as type and size of rating scale, evaluation mode (Section 4.2) etc., are relatively easy to determine from papers, the confusion over which evaluations assess which aspect of quality, and the paucity of detail about experimental design in the great majority of papers, at present mean we do not have a basis for establishing comparability, calling into question the validity of results from reproducibility and meta-evaluation tests that assume comparability.

3 Disentangling Properties of Evaluations

3.1 Similarity of evaluations

When different papers report human evaluations of ***Readability***, we are likely to expect them to report similar system rankings when applied to the same set of system outputs, and similar correlations in meta-evaluations of metrics. But would that expectation change if we then learn that one evaluation measured reading time (on the assumption that more readable texts are faster to read), and in the other, participants were asked to explicitly rate the readability of outputs on a 5-point scale? And what if we are then told that definitions of Readability and questions put to evaluators differed in each case? The point is that we need to know how similar evaluations are, and in what respects, to inform expectations of similarity between their results.

Conversely, when results are reported for different criteria (names), we may expect meta-evaluation and correlation analysis to yield distinguishable results. This can be the case, e.g. Belz and Reiter (2006) report high Pearson correlation with all metrics for Fluency (of weather forecasts), but no correlations with any metrics for Accuracy (of the meteorological information). However, extreme positive correlations ($r = 0.93..0.99$) are often reported (Belz and Kow, 2009; Gardent et al., 2017; Dušek et al., 2020) for pairs of apparently very different quality criteria (e.g. Readability/Meaning Similarity), even when assessed separately for the express purpose of avoiding conflation (Mille et al., 2018, 2019; Dušek et al., 2020).

What is clear, if nothing else, is that some evaluations are less similar, and others more, than meets the eye, and that we do not currently have a systematic way of telling in what respects (in terms of which properties) evaluations are the same and in what respects they are different. In order to be able to do this, we need a system that specifies what those properties are, and provides definitions that make it possible to determine whether evaluations are the same or different in terms of each property.

Identifying such properties is a major challenge, with currently little to no consensus about which ones usefully to distinguish. One of the most basic distinctions is between *what* is being evaluated and *how* it is being evaluated. The former refers to the specific aspect of quality (the *quality criterion*) that an evaluation aims to assess, while the latter refers to how it is mapped to a specific measure that can be implemented in an evaluation experiment. It is worth distinguishing the *how* from the *what*, because in principle there can be many different specific measures and experimental designs that can be used to assess the same quality criterion. Yet the distinction is rarely made in papers, contributing to obscuring similarities between evaluations.

Definitions of *what* is being evaluated often refer to evaluator perception, task success, or preference judgements, all to do with *how* outputs are evaluated. E.g. Allman et al. (2012) define ***Productivity*** as *"the quantity of text an experienced translator could translate in a given period of time [compared] with the quantity of text generated by [the system] that the same person could edit in the given time."* The aspect of quality that is being assessed is the overall quality of a translation given the source text (the better the translation the faster the post-editing) which is measured as the increase in translation speed afforded by use of the system. This is comparable to other assessments of overall translation quality (such as the *Would you use this system* evaluations from dialogue Walker et al. 2001), and results can be expected to be similar, but it is hard to tell this is so, because the required information is not provided in papers.

Properties relating to *how* a quality criterion is evaluated further fall into those that are more 'implementational' in character, such as what type of rating scale is used, with how many possible values, how many evaluators, system outputs, etc., and those can be implemented in different ways such as whether multiple outputs are ranked or single outputs are evaluated separately.

3.2 Disentangled evaluation properties

The proposed system disentangles characteristics of evaluations into 18 properties, each with a set

of possible values, that fall into three groups as indicated above (*quality criteria*, *evaluation mode*, and *experimental design*), and in combination fully specify an evaluation experiment. A **quality criterion** is a criterion in terms of which the quality of system outputs is assessed, and is in itself entirely agnostic about how it is evaluated.

Evaluation modes are properties that need to be specified to turn a quality criterion into an **evaluation measure** that can be implemented, and are orthogonal to quality criteria, i.e. any given quality criterion can be combined with any mode. We distinguish three modes (see Section 4.2).

Experimental design is the full specification of how to obtain a quantitative or qualitative *response value* for a given evaluation measure, yielding a fully specified **evaluation method**. In sum:

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

This three-way separation of properties, and its details in the next section, are motivated by the need to establish comparability in two main contexts: (i) meta-evaluation: comparability assessments of evaluation methods are needed to inform design of meta-evaluation studies and conclusions drawn from them; and (ii) reproducibility testing: similarity in terms of the quality criterion properties indicates which evaluations *should* reproduce each other's results, while similarity in evaluation mode and experimental design can be used to define degrees of reproducibility (Section 6).

4 Classification System

4.1 Quality Criterion properties

The three quality criterion properties are intended to help determine whether or not the same aspect of quality is being evaluated. To this end, we use three properties to characterise quality criteria reflecting (i) what type of quality is being assessed (Section 4.1.1); (ii) what aspect of the system output is being assessed (Section 4.1.2); and (iii) whether system outputs are assessed in their own right or with reference to some system-internal or system-external frame of reference (Section 4.1.3).

4.1.1 Type of quality being assessed

The primary distinction we draw is between criteria assessing correctness, goodness and features. For the former two, it is normally clear which end of the scale is preferred regardless of evaluation context. E.g. one would normally want output texts to be more fluent, more grammatical, more clear.[5]

For feature-type criteria this does not hold; in one evaluation context, one end of the scale might be preferable, in another, the other, and in a third, the criterion may not apply. E.g. when evaluating a conversational agent, *Conversationality* is desirable, but it may not be relevant in a flight booking system. Similarly, *Funny* and *Entertaining* might be desirable properties for a narrative generator, but are inappropriate in a nursing report generator.

We define the three classes as follows:

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* may be 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 Aspect of system output being assessed

Properties in this group capture which aspect of an output is being assessed:

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 aim to assess the content/meaning of the output alone, e.g. *Meaning Preservation* only assesses output

[5] Exceptionally, a goodness/correctness criterion can become a feature, e.g. in expressionist poetry generation where less fluency might be better, as pointed out by a reviewer.

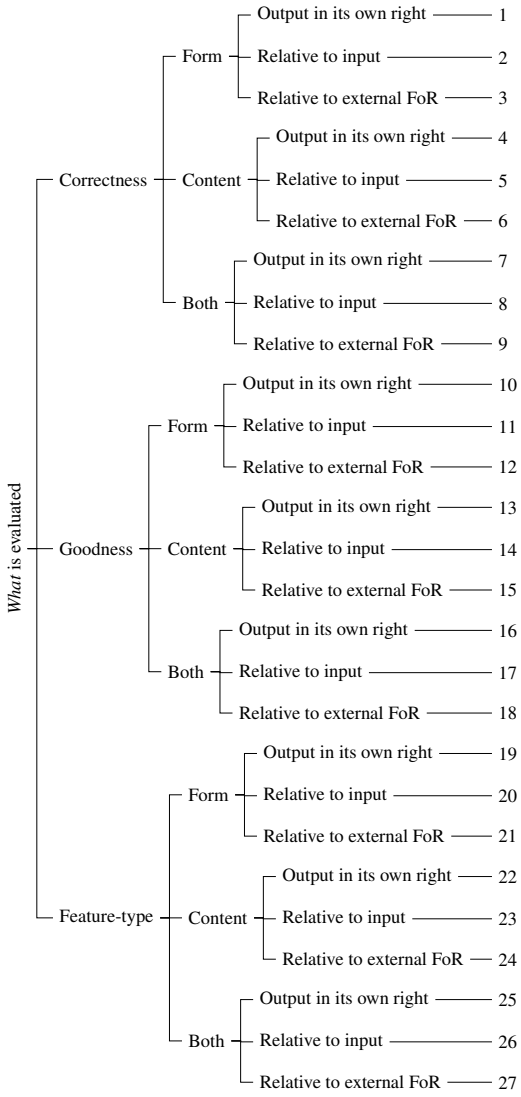

Figure 1: Quality-criterion properties and the 27 different groupings they define (FoR = frame of reference).

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 Quality with/without frame of reference

Properties in this group describe whether assessment of output quality involves a frame of reference in addition to the outputs themselves, i.e. whether the evaluation process also consults (refers to) anything else. We distinguish three cases:

1. **Quality of output in its own right**: assessing output quality 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, e.g. a knowledge base, a person's individual writing style, or an embedding system. E.g. *Factual Accuracy* assesses outputs relative to a source of real-world knowledge.

Figure 1 shows how the quality-criterion properties combine to give 27 groups of quality criteria, numbered for ease of reference in subsequent sections.[6]

4.2 Evaluation Mode Properties

Evaluation modes 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 decidedly more frequent than others). We distinguish three evaluation modes:

1. **Objective vs. subjective**: whether the evaluation involves an objective or a subjective assessment. Examples of objective assessment include any automatically counted or otherwise quantified measurements such as mouse-clicks, occurrences in text, etc. 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.

2. **Absolute vs. relative**: whether evaluators are shown outputs from a single system during evaluation (absolute), or from multiple systems in parallel (relative), in the latter case typically ranking or preference-judging them.

3. **Extrinsic vs. intrinsic**: whether evaluation assesses quality of outputs in terms of theieffect on something external to the system, e.g. performance of an embedding system or of a user at a task (extrinisic), or not (intrinsic).

[6]The tree structure is just a way of showing how the groups relate to each other, we could have used a table instead.

4.3 Experimental Design properties

The properties in this section characterise how response values are obtained for a given evaluation measure defined by quality criterion and evaluation modes. We distinguish 12 properties:

1. System outputs: (1.1) number and (1.2) how selected for inclusion in evaluation.
2. Evaluators: (2.1) number, (2.2) type (expert, how related to the authors, paid/offered course credit/voluntary, etc.), (2.3) how recruited.
3. Method for determining effect size and significance of findings.
4. Scale or other rating instrument: (4.1) size, (4.2) list or range of possible response values, (4.3) how presented to evaluators.
5. Form of response elicitation: e.g. direct quality estimation, (dis)agreement with statement, user-system interaction measurements, etc.
6. Information given to evaluators: Training, instructions, user interface, verbatim question/prompt/etc. seen during evaluation.
7. Experimental conditions: in the lab, in the wild, on crowdsourcing platform, etc.

It is not possible to classify a sample of papers in terms of experimental design properties, because very few provide much of the information. The main relevance of the experimental design properties to the present context is that reproducibility in the narrowest sense (Section 6) assumes that experimental design is the same in the above sense.

5 Example Classifications

Table 1 gives example classifications using the proposed system for evaluation measures[7] from 19 different papers, alongside the criterion name used in the paper. Table 2 shows the corresponding definitions given in the paper (or other evidence if none provided), and maps each evaluation measure to one of the groups from Figure 1. The two tables are divided into five groups, as indicated by the grey text inserts in Table 1. Group 1 contains evaluation measures where the criterion name is the same (***Fluency*** and ***Coherence***, respectively), but quality-criterion properties differ. In conjunction with the definitions in Table 2, this demonstrates that the three evaluation measures called ***Fluency*** in the papers in actual fact assess distinct aspects of quality, as do the four criteria called ***Coherence***. The third example of ***Fluency*** in this group in fact assesses three distinct aspects of quality, which is likely to place a high cognitive load on evaluators.

The three evaluation measures in Group 2 have identical classifications but different names. Based on our classifications and information in the original paper, these criteria are not actually distinct.

The four evaluation measures in Group 3 present a similar case, with ***Reading time*** and ***Ease of reading*** on the one hand, and ***Task success*** and ***Usefulness*** on the other. However, here the evaluation modes are different within each pair.

Group 4 has two examples of feature-type criteria with different names but the same quality-criterion classification; evaluation modes are different, with one involving system rankings (relative mode), and the other direct ratings (absolute mode). The names used (***Text complexity*** and ***Simplicity***) indicate two ends of the same scale, either one of which may be preferable depending on context.

The evaluation measures in Group 5 involve quality criteria that appear at first glance closely related (see Table 2). What they have in common is that they assess aspects of the quality of referring expressions. However, none of the classifications are exactly the same, and we would argue that the criteria assess distinct aspects of quality: correct pronoun usage, identifiability of referents, and fast referent identification, the former two being correctness criteria, the latter a goodness criterion.

6 Use Cases

The proposed classification system provides a basis for systematically comparing evaluation methods. We can see at least three contexts in which this is either a prerequisite or at least useful, as outlined in the next three subsections.

6.1 Design and reporting of evaluations

At present, in the majority of cases it is generally not clear enough from papers what quality criterion was evaluated in a human evaluation, one of the main conclusions we drew from our attempt to map quality criteria reported in papers to normalised terms and definitions in our extensive survey of human evaluation in NLG (Howcroft et al., 2020). The example classifications we give in Tables 1 and 2 represent our interpretation of the information provided in each paper, but the authors may have intended slightly different meanings, e.g. for

[7] We're not including experimental design properties for reasons explained in the preceding section.

Paper	Criterion Name in Paper	quality-criterion properties			Evaluation Mode		
		Type of Quality	Form/ Content	Frame of Reference (FoR)	obj. / subj.	abs. / rel.	extr. / intr.
Group 1 – Same name, different quality-criterion properties, same evaluation modes (2 example sets):							
Yu et al. (2020)	*Fluency*	goodness	form	none	subj.	abs.	intr.
Van de Cruys (2020)	*Fluency*	correctness	form	none	subj.	abs.	intr.
Pan et al. (2020)	*Fluency*	correctness	(a) form (b) content (c) content	(a) none (b) none (c) external FoR	subj.	abs.	intr.
Van de Cruys (2020)	*Coherence*	goodness	content	none	subj.	abs.	intr.
Juraska et al. (2019)	*Coherence*	(a) correctness (b) goodness	form	none	subj.	abs.	intr.
Chai and Wan (2020)	*Coherence*	goodness	content	external FoR	subj.	abs.	intr.
Barros et al. (2017)	*Coherence*	correctness	content	none	subj.	abs.	intr.
Group 2 – Different names, same quality-criterion properties, same evaluation modes:							
Wang et al. (2020)	*Faithfulness*	correctness	content	FoR = input	obj.	abs.	intr.
Cao et al. (2020)	*Content Similarity*	correctness	content	FoR = input	obj.	abs.	intr.
Zhou et al. (2020)	*Content Preservation*	correctness	content	FoR = input	obj.	abs.	intr.
Group 3 – Different names, same quality-criterion properties, different evaluation modes (2 example sets):							
Gatt and Belz (2008)	*Reading Time*	goodness	both	none	obj	abs	extr
Forrest et al. (2018)	*Ease of Reading*	goodness	both	none	subj.	abs.	intr.
Miliaev et al. (2003)	*Usefulness*	goodness	both	external FoR	subj.	abs.	intr.
Qu and Green (2002)	*Task success*	goodness	both	external FoR	obj.	abs.	extr.
Group 4 – Equivalent names, same quality-criterion properties, different evaluation modes:							
Moraes et al. (2016)	*Text Complexity*	feature	both	none	subj.	rel.	intr.
Narayan and Gardent (2016)	*Simplicity*	feature	both	none	subj.	abs.	intr.
Group 5 – Different names, different quality-criterion properties, different evaluation modes, related definitions:							
Chai and Wan (2020)	*Coreference*	correctness	both	none	subj.	abs.	intr.
Funakoshi et al. (2004)	*Accuracy*	correctness	both	external FoR	obj.	abs.	extr.
Gatt and Belz (2008)	*Identification Time*	goodness	both	external FoR	obj	abs	extr

Table 1: Examples of human evaluations described according to the proposed classification system.

Fluency some authors might take that to relate to both form and content. As things stand, it is often impossible to tell, because (a) there is not enough information provided in papers, and (b) even if there is, it is not described in shared terms.

A related question is how well evaluators understand what they are being asked to evaluate. It is often assumed that aspects of quality like Fluency and Clarity, and the differences between them, are intuitively clear to evaluators, but how certain is this when good intra and inter-evaluator agreement is so hard to achieve (Belz and Kow, 2011), and correlations between apparently very different criteria are so often in the high nineties (Section 3)? That researchers struggle to explain what to evaluate is also clear from definitions and prompts reported in papers which often define one quality criterion in terms of others (e.g. Rows 2, 3, 5 in Tables 1 and 2), and use inconsistent language in quality criterion name, definition, and prompts.

A shared classification system helps address both the above, (a) making clear what needs to be included in reports to convey what was evaluated, and (b) providing a basis for conveying to evaluators what aspect of quality they are expected to assess in such a way as to ensure multiple evaluators end up with the same interpretation as each other and as the designers of the experiment.

6.2 Meta-evaluation

The standard way of validating a new automatic evaluation metric is to obtain system-level correlations with human assessments of the same set of system outputs, usually termed meta-evaluation. The expectation that a given metric *should* correlate with the human evaluation it is meta-evaluated against is not normally justified, but the implicit assumption is that they are measuring the same thing, for why else should they correlate?

For example, years of mixed results from meta-evaluating BLEU against a wide variety of different human evaluations have resulted in conclusions that BLEU is not a good metric, or is not reliable enough, because it does not correlate consistently well with human evaluations. But why should a single metric be expected to correlate equally well with human assessments of quality criteria as distinct as Fluency and Accuracy (of content)?

Paper	Criterion Name in Paper	Definition/evidence	Suggested class	
			#	gloss
Yu et al. (2020)	*Fluency*	"judging the question fluency"	10	goodness of form of output iioR
Van de Cruys (2020)	*Fluency*	"is the poem grammatical and syntactically well-formed?"	1	correctness of form of output iioR)
Pan et al. (2020)	*Fluency*	"indicates whether the question follows the grammar and accords with the correct logic"	1	(a) correctness of form of output iioR
			4	(b) correctness of content of output iioR
			6	(c) correctness of content relative to ext. FoR
Van de Cruys (2020)	*Coherence*	"[whether] the poem [is] thematically structured"	13	goodness of content of output iioR
Juraska et al. (2019)	*Coherence*	"measures [a conversation's] grammaticality and fluency"	1	(a) correctness of form of output iioR
			10	(b) goodness of form of output iioR
Chai and Wan (2020)	*Coherence*	"measures if a question is coherent with previous ones"	15	goodness of content relative to ext. FoR
Barros et al. (2017)	*Coherence*	"meaning of the generated sentence, [...] sentence with no meaning would be rated with 1 and a sentence with a full meaning would be rated with 5"	4	correctness of content of output iioR
Wang et al. (2020)	*Faithfulness*	"A sentence is faithful if it contains only information supported by the table. [...] Also, the generated sentence should cover as much information in the given table as possible."	5	correctness of content relative to input
Cao et al. (2020)	*Content Similarity*	"measures how much content is preserved during style transfer"	5	correctness of content relative to input
Zhou et al. (2020)	*Content Preservation*	"preservation of original content"	5	correctness of content relative to input
Gatt and Belz (2008)	*Reading Time*	"[time] from the point at which the description was presented, to the point at which a participant called up the next screen via mouse click"	16	goodness of form/content of output iioR
Forrest et al. (2018)	*Ease of Reading*	"self-reported ease of reading of the explanation and interpretation"	16	goodness of form/content of output iioR
Miliaev et al. (2003)	*Usefulness*	'how useful was the manual to cope with the task"	18	goodness of form/content relative to ext. FoR
Qu and Green (2002)	*Task success*	"the degree of task success with respect to the user's original information need"	18	goodness of form/content relative to ext. FoR
Moraes et al. (2016)	*Text Complexity*	"ability of the system on varying the text complexity as perceived by human readers."	25	complexity of form/content of output iioR
Narayan and Gardent (2016)	*Simplicity*	"How much does the generated sentence(s) simplify the complex input?"	25	complexity of form/content of output iioR
Chai and Wan (2020)	*Coreference*	"measures if a question uses correct pronouns"	7	correctness of form/content of output iioR
Funakoshi et al. (2004)	*Accuracy*	"rates at which subjects could identify the correct target objects from the given expressions"	9	correctness of form/content relative to ext. FoR
Gatt and Belz (2008)	*Identification Time*	"[time] from the point at which pictures [...] were presented on the screen to the point where a participant identified a referent by clicking on it"	18	goodness of form/content relative to ext. FoR

Table 2: Companion table to Table 1. Definitions/other evidence from each paper, suggested mapping to groups from Figure 1, and gloss for each group (FoR = frame of reference; iioR = in its own right).

Even for conclusions about correlation with human assessments of individual quality criteria, such as that BLEU does not correlate consistently well with Fluency, the implicit assumption is that all evaluations assessing something called 'Fluency' in fact succeed in measuring the same thing. Looking at the first three rows of Tables 1 and 2 it is doubtful that we currently know whether or not BLEU does correlate consistently with Fluency.

A shared classification system for human evaluation methods provides firmer ground for conclusions by helping establish which groups of human evaluations are similar enough to be expected to correlate similarly with a given metric, and even whether a given metric is similar enough to a given type of human evaluation to be expected to corre-

late well with it.

6.3 Reproducibility tests

In simple terms, reproducibility tests re-run existing evaluations in either the same way or with controlled differences to see if the results are the same. Beyond this, there is little agreement in NLP/ML, despite growing levels of interest in the subject of reproducibility over recent years. Not wishing to wade into the general debate, we use the definitions of the International Vocabulary of Metrology (VIM) (JCGM, 2012), where **repeatability** is the precision of measurements of the same or similar object obtained under the same conditions, as captured by a specified set of **repeatability conditions**, whereas **reproducibility** is the precision of measurements of the same or similar object obtained under different conditions, as captured by a specified set of **reproducibility conditions**.[8]

The properties defined by the classification system proposed here can be straightforwardly used to serve as the set of repeatability/reproducibility conditions, for repeatability specifying the respects in which original and repeat measurements are controlled to be the same, and for reproducibility additionally specifying in which respects original and reproduction measurements differ.

One step further would be to select nested subsets of properties to define different *degrees* of reproducibility, for example:

1. Reproducibility in the first degree: all 18 properties are the same.
2. Reproducibility in the second degree: quality criteria properties and evaluation mode properties are the same, but some or all of the experimental design properties differ.
3. Reproducibility in the third degree: quality criteria properties are the same, but some or all of the evaluation mode properties and experimental design properties differ.

Such degrees of reproducibility are similar in spirit to the four-way 'quadrants of reproducibility' proposed recently by Whitaker (2017) and adopted by Schloss (2018), but unlike them, the above approach (a) is not inherently limited to just two dimensions (data and code), and (b) does not attach disputed labels (replicability, robustness, generalisability) to the different degrees.

[8]The ACM definitions are described as being based on VIM but it's not clear how exactly: https://www.acm.org/publications/policies/artifact-review-and-badging-current

7 Future Work and Conclusions

The present paper is intended as a step towards full comparability of human evaluation methods in NLG. There are clear directions for further development. E.g. we have remained agnostic about what happens within the 27 groups of quality criteria defined by the proposed system (visualised in Figure 1). Do the groups map to 27 quality criteria that are enough for all evaluation contexts, merely needing to be 'localised' to a specific task and domain? This might work for correctness criteria and goodness criteria, but new criteria can be almost arbitrarily added to the feature-type groups.

Another question is how to ensure that experimental design matches a chosen quality criterion and does not end up evaluating something else entirely. We have pointed to using the terms and definitions of the proposed classification properties in experimental design, but not given details of how this can be done. We can see relevance also to recent machine-learned evaluation metrics (to clarify what it is they are emulating). We plan to address the above lines of inquiry in future work.

While this paper proposes a standard way of classifying evaluation methods, we do not propose a standardised nomenclature of quality criterion names and definitions. If such a standard did become widely adopted in the field, it would go a long way towards addressing the issue of comparability. However, given the deeply ingrained habit in NLG of using ad-hoc, tailored evaluation methods that differ widely even within small NLG subfields, this seems unrealistic for now.

Our aim in this paper has instead been to find a way of teasing apart the similarities and dissimilarities of evaluation methods used in the current, highly diverse context, to yield a set of clearly defined properties that provides a firm basis for designing and reporting evaluation methods, establishing comparability for meta-evaluation, and specifying repeatability/reproducibility conditions for reproducibility tests.

8 Acknowledgements

We thank our reviewers for their valuable feedback. Mille's contribution was supported by the European Commission under the H2020 contracts 870930-RIA, 779962-RIA, 825079-RIA, 786731-RIA.

References

Tod Allman, Stephen Beale, and Richard Denton. 2012. Linguist's assistant: A multi-lingual natural language generator based on linguistic universals, typologies, and primitives. In *INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference*, pages 59–66, Utica, IL. Association for Computational Linguistics.

Cristina Barros, Dimitra Gkatzia, and Elena Lloret. 2017. Improving the naturalness and expressivity of language generation for Spanish. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 41–50, Santiago de Compostela, Spain. Association for Computational Linguistics.

Anya Belz and Eric Kow. 2009. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 16–24.

Anya Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*.

Anya Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 230–235, Portland, Oregon, USA. Association for Computational Linguistics.

Anya Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Zi Chai and Xiaojun Wan. 2020. Learning to ask more: Semi-autoregressive sequential question generation under dual-graph interaction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 225–237, Online. Association for Computational Linguistics.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, 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.

Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*, pages 54–63, Philadelphia, Pennsylvania, U.S.A. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123–156.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech Language*, 59:123 – 156.

James Forrest, Somayajulu Sripada, Wei Pang, and George Coghill. 2018. Towards making NLG a voice for interpretable machine learning. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 177–182, Tilburg University, The Netherlands. Association for Computational Linguistics.

Kotaro Funakoshi, Satoru Watanabe, Naoko Kuriyama, and Takenobu Tokunaga. 2004. Generating referring expressions using perceptual groups. In *International Conference on Natural Language Generation*, pages 51–60, Brockenhurst, UK. Springer.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from rdf data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133.

Albert Gatt and Anya Belz. 2008. Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 50–58, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Vrindavan Harrison and Marilyn Walker. 2018. Neural generation of diverse questions using answer focus, contextual and linguistic features. In *Proceedings of the 11th International Conference on Natural*

Language Generation, pages 296–306, Tilburg University, The Netherlands. 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*.

Joint Committee for Guides in Metrology, JCGM. 2012. International vocabulary of metrology: Basic and general concepts and associated terms (VIM).

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Chris van der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the wmt19 metrics shared task: Segment-level and strong mt systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90.

Joy Mahapatra, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2016. Statistical natural language generation from tabular non-textual data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 143–152, Edinburgh, UK. Association for Computational Linguistics.

Nitika Mathur, Tim Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *arXiv preprint arXiv:2006.06264*.

Jonathan May and Jay Priyadarshi. 2017. SemEval-2017 task 9: Abstract Meaning Representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545, Vancouver, Canada. Association for Computational Linguistics.

Nestor Miliaev, Alison Cawsey, and Greg Michaelson. 2003. Applied NLG system evaluation: FlexyCAT. In *Proceedings of the 9th European Workshop on Natural Language Generation (ENLG-2003) at EACL 2003*, pages 55–62, Sofia, Bulgaria.

Simon Mille, Anya Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the 1st Workshop on Multilingual Surface Realisation (MSR), 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–12, Melbourne, Australia.

Simon Mille, Anya Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The second multilingual surface realisation shared task (SR'19): Overview and evaluation results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. Association for Computational Linguistics.

Priscilla Moraes, Kathleen Mccoy, and Sandra Carberry. 2016. Enabling text readability awareness during the micro planning phase of NLG applications. In *Proceedings of the 9th International Natural Language Generation conference*, pages 121–131, Edinburgh, UK. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 105–113, Dublin, Ireland.

Shashi Narayan and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Edinburgh, UK. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Information Processing Management*, 43(6):1506 – 1520. Text Summarization.

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.

Yan Qu and Nancy Green. 2002. A constraint-based approach for cooperative information-seeking dialogue. In *Proceedings of the International Natural Language Generation Conference*, pages 136–143, Harriman, New York, USA. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ehud Reiter and Anya Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

Sashank Santhanam and Samira Shaikh. 2019. Towards best experiment design for evaluating dialogue system output. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 88–94, Tokyo, Japan. Association for Computational Linguistics.

Patrick D Schloss. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3).

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. 2018. WebNLG challenge: Human evaluation results. Technical report, Technical report.

Sebastian Varges. 2006. Overgeneration and ranking for spoken dialogue systems. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 20–22, Sydney, Australia. Association for Computational Linguistics.

Marilyn Walker, J Aberdeen, J Boland, E Bratt, J Garofolo, Lynette Hirschman, A Le, Sungbok Lee, Shrikanth Narayanan, Kishore Papineni, et al. 2001. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Seventh European Conference on Speech Communication and Technology*.

Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1072–1086, Online. Association for Computational Linguistics.

Kirstie Whitaker. 2017. Publishing a reproducible paper. Presentation at Open Science in Practice Summer School, `https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf`.

Sander Wubben, Emiel Krahmer, Antal van den Bosch, and Suzan Verberne. 2016. Abstractive compression of captions with attentive recurrent neural networks. In *Proceedings of the 9th International Natural Language Generation conference*, pages 41–50, Edinburgh, UK. Association for Computational Linguistics.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75, Tokyo, Japan. Association for Computational Linguistics.

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.

Hongyu Zang and Xiaojun Wan. 2017. Towards automatic generation of product reviews from aspect-sentiment scores. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 168–177, Santiago de Compostela, Spain. Association for Computational Linguistics.

Chulun Zhou, Liangyu Chen, Jiachen Liu, Xinyan Xiao, Jinsong Su, Sheng Guo, and Hua Wu. 2020. Exploring contextual word-level style relevance for unsupervised style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7135–7144, Online. Association for Computational Linguistics.