Arguing for consistency in human evaluation of natural language generation systems

David M. Howcroft

EDINBURGH CENTRE FOR ROBOTICS

HERIOT WATT UNIVERSITY

The Problem

There is **no common scale** or format **for text quality**.

Why is that a problem?

Lack of clarity about what we care about

- sampling participants' notions of fluency, etc?
- 'objective' grammaticality?
- overall quality vs. features correlated with quality?

Challenges to comparability

- inconsistent reporting
- defining terms for subjects?
- providing training items?
- examining how participants use our scales?
- inconsistent statistical analysis

Evaluating 'Fluency'

'linguistic quality of the text' (Gatt & Krahmer 2018)

But which quality do we want to assess?

Clarity

'how clear the description is'
'The message of this text is completely clear to me.'

Fluency

'how fluent a sentence is on its own'
'How do you judge the fluency of Sentence B?'

Grammaticality

'How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?'
'How would you grade the syntactic quality of the [text]?'
'This text is written in proper Dutch.'

Naturalness

'Could the utterance have been produced by a native speaker?'

Readability

'How hard was it to read the [text]?'
'This is sometimes called "fluency", and ... decide how well the highlighted sentence reads; is it good fluent English, or does it have grammatical errors, awkward constructions, etc.'
'This text is easily readable.'

Understandability

'How easy is this [text] to understand?'
'How clear (easy to understand) is the highlighted sentence within the context of the text extract?'
'The ... summary was easy to understand.'
'Which system's responses were easier to understand?'
'While reading, I immediately understood the text.'

Structure of an evaluation

Instructions

- just evaluation? (intrinsic)
- task completion? (extrinsic)
- *Do we keep participants naive?*

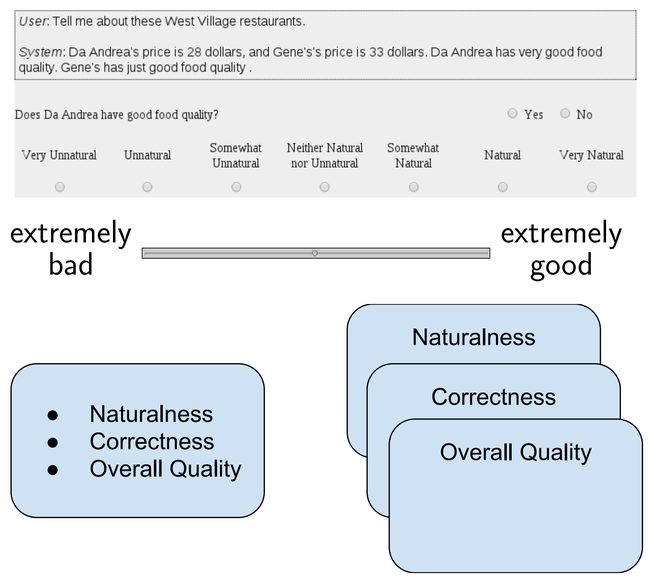
Definitions

- provided?
- always accessible?

Training

- provided? validated? with feedback?

Questions vs. prompts
Discrete vs. continuous
Scored vs. ranked
Independent vs. joint



Evaluating 'Adequacy'

'**accuracy, adequacy, relevance or correctness** relative to the input' (Gatt & Krahmer 2018)

Accuracy or Completeness

'How much of the meaning expressed in [text] A is also expressed in [text] B?'
'The ... summary is accurate.'
'Does the utterance provide all the useful information from the meaning representation?'
'Did you find all the information you were looking for?'

Informativeness or Relevance

'Would it be *correct* to convey this information by saying...?'
'Did the help provide you with enough information...?'
'which system offered you more information'
'Was the information provided ... relevant for your task'?

- correctness, similarity, truthfulness, importance, meaning preservation, non-redundancy, not misleading, sensibility...

What can we do?

- Guidelines for reporting (cf. van der Lee et al. 2019)
- Agree on definitions of these terms
- Build templates for evaluation
- Systematically approach statistics **our own ongoing work!**

Recommended Reading

Amidei, Piwek, & Willis. 2018. 'Evaluation methodologies in Automatic Question Generation 2013-2018'. *INLG*.
Amidei, Piwek, & Willis. 2019. 'The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations'. *INLG*.
Belz & Kow. 2011. 'Discrete vs. Continuous Rating Scales for ... Evaluation in NLP'. *ACL*.
Gatt & Krahmer. 2018. 'Survey of the State of the Art in Natural Language Generation: Core tasks, applications, and evaluation'. *JAIR*.
Gkatzia & Mahamood. 2015. 'A Snapshot of NLG Evaluation Practices 2005–2014.' *ENLG*. Novikova, Dušek, Cercas Curry, & Rieser. 2017. 'Why We Need New Evaluation Metrics for NLG'. *EMNLP*.
Novikova, Dušek, & Rieser. 2018. 'RankME: Reliable Human Ratings for NLG'. *NAACL*.
van der Lee, Gatt, van Miltenburg, Wubben, & Krahmer. 2019. 'Best practices for the human evaluation of automatically generated text'. *INLG*.

D.Howcroft@hw.ac.uk
Heriot-Watt University