Breaking News

It’s time to fix evaluation of generated text

Sebastian Gehrmann
gehrmann@google.com
@SebGehr
Google Research
Good News! We are making progress toward solving summarization.

https://paperswithcode.com/sota/document-summarization-on-cnn-daily-mail?metric=ROUGE-L

Sebastian Gehrmann, Google Research, 2021
Good News! We are making progress toward solving summarization.
Reference Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the Seattle band Silkworm.

Candidates

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the California band Grateful Dead.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer from Seattle, Washington.
Reference Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the Seattle band Silkworm.

Candidates

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer in the California band Grateful Dead.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer.

Michael Dahlquist (December 22, 1965 - July 14, 2005) was a drummer from Seattle, Washington.

BLEU	ROUGE

0.79	0.77

0.71	0.79

0.73	0.70

Metrics prefer bad generations over good ones.

Dhingra et al., 2019
ROUGE and its problems...

“ROUGE may not be a good method for measuring the usefulness of summaries when the summaries are not extractive.”
Dorr et al., 2005

A system’s ability to produce human-like outputs may be completely unrelated to its effect on human task-performance.
Belz+Gatt, 2008

Metrics may provide a useful measure of language quality, although the evidence for this is not as strong as we would ideally like to see; however, they do not provide a useful measure of content quality.
Reiter+Belz, 2009

Luckily, we are not using ROUGE to measure content quality of abstractive summaries, right? ....... Right?
Agenda

01 A brief preview
02 Automatic evaluation is broken
03 Human evaluation is broken
04 Datasets are broken
05 How do we fix things?
Automatic Evaluation is broken
We suspect that ROUGE is not great.

So let’s see what people are using.
Summarization is dominated by ROUGE-1, -2, and -L.
Fun fact: The selection was popularized by Rush et al. (2015), who picked a subset of the DUC-2004 options which also included 3, 4, and LW. But ROUGE-2 and ROUGE-SU4 were used in later DUC challenges.
A very scientific survey.

I read 20 modeling-focused summarization papers from ACL 2021 and recorded the following evaluation aspects:

1) Automatic metrics
2) Human evaluation criteria [if applicable]
3) Dataset(s)

Throughout the talk, I will show the results.
A very scientific survey.

I read 20 modeling-focused summarization papers from ACL 2021 and recorded the following evaluation aspects:

1) Automatic metrics
2) Human evaluation criteria [if applicable]
3) Dataset(s)

Throughout the talk, I will show the results. On the right, you can see the metrics.
But just how bad is ROUGE?

And how do you evaluate a metric?
Let’s look at some more recent studies.
ROUGE F-Scores may not be enough.

Recall, that the use of F-scores for ROUGE 1, 2, and L is essentially arbitrary. It may also be strictly suboptimal.

In a study correlation of assessment scores of all possible 192 ROUGE configurations found that the best performing one was to use BLEU instead.¹

The best ROUGE was ROUGE-2 precision with stemming and removed stopwords.

→ If using ROUGE, consider reporting fine-grained scores.

¹Not statistically significant, though.
How can we evaluate with lexical overlap, if humans don’t even agree with each other?

Data: 100 samples from the CNN/DM test set.

Unconstrained: Every annotator selects sentences in the input they consider important.

Constrained: Every annotator selects sentences with answers to three questions related to the document.

Even when only 3/5 people have to agree on a sentence, there is 0.6 sentences per document on which all agree.

→ When there is only one reference, we can’t use lexical overlap to capture everyone’s summarization preferences.

<table>
<thead>
<tr>
<th>Human vote threshold</th>
<th>Sent. per article considered important</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unconstrained</td>
</tr>
<tr>
<td>= 5</td>
<td>0.028</td>
</tr>
<tr>
<td>≥ 4</td>
<td>0.213</td>
</tr>
<tr>
<td>≥ 3</td>
<td>0.627</td>
</tr>
<tr>
<td>≥ 2</td>
<td>1.695</td>
</tr>
<tr>
<td>≥ 1</td>
<td>5.413</td>
</tr>
</tbody>
</table>

Kryscinski et al., 2019
The correlation between human judgements and ROUGE is poor.

For 100 CNN/DM test examples, ask 5 raters to judge:

- **Relevance**: selection of important content from the source
- **Consistency**: factual alignment between the summary and the source
- **Fluency**: quality of individual sentences
- **Coherence**: collective quality of all sentences.

All 5 judgements are averaged.

Then, measure Pearson’s correlation coefficients and Kendall rank correlation coefficients between judgements and ROUGE.

<table>
<thead>
<tr>
<th></th>
<th>Pearson correlation</th>
<th>Kendall rank correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-1</td>
<td>R-2</td>
</tr>
<tr>
<td>Relevance</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Consistency</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Fluency</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Coherence</td>
<td>0.05</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Kryscinski et al., 2019
Repeating the evaluation at scale does not result in (much) better results

Same dataset + criteria, but 8 annotations per example (5 Mturkers, 3 experts) and 16 systems.

Results are similarly not great.

Some “unconventional” summarization metrics like ROUGE-3 and METEOR perform better than “standard” ROUGE settings.

→ We need better metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Coherence</th>
<th>Consistency</th>
<th>Fluency</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>0.2500</td>
<td>0.5294</td>
<td>0.5240</td>
<td>0.4118</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>0.1618</td>
<td>0.5882</td>
<td>0.4797</td>
<td>0.2941</td>
</tr>
<tr>
<td>ROUGE-3</td>
<td>0.2206</td>
<td>0.7059</td>
<td>0.5092</td>
<td>0.3529</td>
</tr>
<tr>
<td>ROUGE-4</td>
<td>0.3088</td>
<td>0.5882</td>
<td>0.5535</td>
<td>0.4118</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.0735</td>
<td>0.1471</td>
<td>0.2583</td>
<td>0.2353</td>
</tr>
<tr>
<td>BertScore-p</td>
<td>0.0588</td>
<td>-0.1912</td>
<td>0.0074</td>
<td>0.1618</td>
</tr>
<tr>
<td>BertScore-r</td>
<td>0.1471</td>
<td>0.6618</td>
<td>0.4945</td>
<td>0.3088</td>
</tr>
<tr>
<td>BertScore-f</td>
<td>0.2059</td>
<td>0.0441</td>
<td>0.2435</td>
<td>0.4265</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.1176</td>
<td>0.0735</td>
<td>0.3321</td>
<td>0.2206</td>
</tr>
<tr>
<td>CHRF</td>
<td>0.3971</td>
<td>0.5294</td>
<td>0.4649</td>
<td>0.5882</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.1176</td>
<td>-0.1912</td>
<td>-0.0221</td>
<td>0.1912</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.2353</td>
<td>0.6324</td>
<td>0.6126</td>
<td>0.4265</td>
</tr>
</tbody>
</table>

Kendall-Tau rank correlation of different metrics
A single metric is not enough.

Human annotations of DUC-2004 → almost no correlation between linguistic quality and coverage, but coverage is almost never higher than linguistic quality.

This finding is consistent with Pitler et al. (2010) who find correlations between some evaluation categories, but not between linguistic and content quality.

→ We cannot rely on a single metric to provide all details.

Graham, 2015
We can go deeper. Can we audit metrics?

If we know common hallucinations, we can inject them into the references and test if a metric score decreases.

Scores of a well-calibrated metric should negatively correlate with monotonically increasing number of errors.

Gabriel et al., 2021
We can go deeper. Can we audit metrics?

If we know common hallucinations, we can inject them into the references and test if a metric score decreases. Scores of a well-calibrated metric should negatively correlate with monotonically increasing number of errors.

→ Most metrics are calibrated, but R-1+R-L fail completely.

Also note that this is system-level correlation, not segment.

<table>
<thead>
<tr>
<th></th>
<th>R-1</th>
<th>R-2</th>
<th>R-3</th>
<th>R-L</th>
<th>BERTScore</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upper Bound</strong></td>
<td>10.61</td>
<td>2.56</td>
<td>0.72</td>
<td>9.32</td>
<td>83.76</td>
</tr>
<tr>
<td><strong>Level 1</strong></td>
<td>10.49 / 10.76</td>
<td>2.54 / 2.56</td>
<td>0.70</td>
<td>9.22 / 9.42</td>
<td>83.53 / 83.56</td>
</tr>
<tr>
<td><strong>Level 2</strong></td>
<td>10.40 / 10.86</td>
<td>2.51 / 2.54</td>
<td>0.69 / 0.68</td>
<td>9.16 / 9.49</td>
<td>83.36 / 83.38</td>
</tr>
<tr>
<td><strong>Level 3</strong></td>
<td>10.33 / 10.92</td>
<td>2.49 / 2.52</td>
<td>0.69 / 0.67</td>
<td>9.10 / 9.55</td>
<td>83.21 / 83.26</td>
</tr>
<tr>
<td><strong>Lower Bound</strong></td>
<td>5.44</td>
<td>0.39</td>
<td>0.01</td>
<td>4.94</td>
<td>80.08</td>
</tr>
</tbody>
</table>

| **Correlation**| -1.00 / 0.98 | -0.97 / -1.00 | -0.87 / -1.00 | -1.00 / 1.00 | -1.00 |
| **p-value**     | 0.03* / 0.10 | 0.16 / 0.05*  | 0.33 / 0.05*  | <0.01** / 0.02* | 0.02* / 0.06 |

Left: entity errors, Right: non-entity errors

Gabriel et al., 2021

Sebastian Gehrmann, Google Research, 2021
Let’s look deeper into faithfulness.

A model not faithful if it hallucinates.

**Intrinsic**: A model misrepresents facts in the input
“Former London mayoral candidate” → “Former London mayor”

**Extrinsic**: A model ignores the input
“mayoral candidate Peter” → “mayor Sara”

**Factual**: A model hallucinates facts that are true
“mayoral candidate Peter” → “2016 mayoral candidate Peter”

Factual hallucinations may be acceptable.

Semantic or lexical similarity does not help for these fine-grained determinations.

→ When assessing a model, entailment-type metrics may be necessary to detect hallucinations.

Maynez, Narayan, et al., 2020

<table>
<thead>
<tr>
<th>Metric</th>
<th>Faithful</th>
<th>Factual</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1</td>
<td>0.197</td>
<td>0.125</td>
</tr>
<tr>
<td>ROUGE-2</td>
<td>0.162</td>
<td>0.095</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.162</td>
<td>0.113</td>
</tr>
<tr>
<td>BERTScore</td>
<td>0.190</td>
<td>0.116</td>
</tr>
<tr>
<td>QA</td>
<td>0.044</td>
<td>0.027</td>
</tr>
<tr>
<td>Entailment</td>
<td><strong>0.431</strong></td>
<td><strong>0.264</strong></td>
</tr>
</tbody>
</table>

Table 4: Spearman’s correlation coefficient ($|r_s|$) of different metrics with faithful and factual annotations.
A glimmer of hope on the horizon

Trained metrics can have much higher correlations.¹

“Only” requirements:
1) Many high-quality annotations
2) Large pretrained models

Pu et al., 2021

¹At least in MT
Let’s build better metrics!

But, how do we get people to adopt it? It has to be fast, and easy to use, and work for all languages, and all tasks, and ...

🤗

Great, let’s do that!
Not so fast! We need high-quality data first.

How do we get that?
Human evaluation.

Do we know how to do human eval?
No, not really.

But you said that human eval is necessary! Let’s move to the next section.
Human Evaluation is broken
Takeaways so far

One number cannot characterize all performance aspects of a model output
→ We need **multiple specialized metrics**.

None of our metrics correlate well with human judgements
→ **Human evaluation is a necessary** component of model evaluation.

Trained metrics can potentially have much better correlations
→ We need **many** high quality human annotations.
Coming back to the survey

9/20 papers used human evaluation.

But what were they assessing? 👉

Wide range of criteria, there is no agreement here.

And the problem runs even deeper.

Informativeness 5
Conciseness/Succinctness 4
Fluency 4
Relevance/Salience 4
Coherence 2
Consistency 2
Coverage 1
Error Classifications 1
Factualness 1
Faithfulness 1
Grammaticality 1
Meaning-Preserving 1
What is being measured?

In a study of 478 INLG papers, the authors found:
- 204 unique names of quality criteria.
- 71 truly different aspects.

Similar aspects may be considered equal by readers: Spelling Accuracy vs. Correctness of the Surface Form.

Often, details are not provided:
- >50% missing definitions (279/478)
- ~66% missing evaluator prompts/questions (311/478)
- 20% missing criteria names (98/478)

We need to understand what is measured and not group different annotations into one bucket.

Table 4: Occurrence counts for normalised criterion names.

<table>
<thead>
<tr>
<th>Criterion Paraphrase</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>usefulness for task/information need</td>
<td>39</td>
</tr>
<tr>
<td>grammaticality</td>
<td>39</td>
</tr>
<tr>
<td>quality of outputs</td>
<td>35</td>
</tr>
<tr>
<td>understandability</td>
<td>30</td>
</tr>
<tr>
<td>correctness of outputs relative to input (content)</td>
<td>29</td>
</tr>
<tr>
<td>goodness of outputs relative to input (content)</td>
<td>27</td>
</tr>
<tr>
<td>clarity</td>
<td>17</td>
</tr>
<tr>
<td>fluency</td>
<td>17</td>
</tr>
<tr>
<td>goodness of outputs in their own right</td>
<td>14</td>
</tr>
<tr>
<td>readability</td>
<td>14</td>
</tr>
<tr>
<td>information content of outputs</td>
<td>14</td>
</tr>
<tr>
<td>goodness of outputs in their own right (both form and content)</td>
<td>13</td>
</tr>
<tr>
<td>referent resolvability</td>
<td>11</td>
</tr>
<tr>
<td>usefulness (nonspecific)</td>
<td>11</td>
</tr>
<tr>
<td>appropriateness (content)</td>
<td>10</td>
</tr>
<tr>
<td>naturalness</td>
<td>10</td>
</tr>
<tr>
<td>user satisfaction</td>
<td>10</td>
</tr>
<tr>
<td>wellorderedness</td>
<td>10</td>
</tr>
<tr>
<td>correctness of outputs in their own right (form)</td>
<td>9</td>
</tr>
<tr>
<td>correctness of outputs relative to external frame of reference (content)</td>
<td>8</td>
</tr>
<tr>
<td>ease of communication</td>
<td>7</td>
</tr>
<tr>
<td>humanlikeness</td>
<td>7</td>
</tr>
<tr>
<td>appropriateness</td>
<td>6</td>
</tr>
<tr>
<td>understandability</td>
<td>6</td>
</tr>
<tr>
<td>nonredundancy (content)</td>
<td>6</td>
</tr>
<tr>
<td>goodness of outputs relative to system use</td>
<td>5</td>
</tr>
<tr>
<td>appropriateness (both form and content)</td>
<td>5</td>
</tr>
</tbody>
</table>
How is it being measured?

<table>
<thead>
<tr>
<th>Form</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct quality estimation</td>
<td>207</td>
</tr>
<tr>
<td>relative quality estimation</td>
<td>72</td>
</tr>
<tr>
<td>(dis)agreement with quality statement</td>
<td>48</td>
</tr>
<tr>
<td>classification</td>
<td>38</td>
</tr>
<tr>
<td>task performance measurements</td>
<td>35</td>
</tr>
<tr>
<td>qualitative feedback</td>
<td>20</td>
</tr>
<tr>
<td>evaluation through post-editing/annotation</td>
<td>18</td>
</tr>
<tr>
<td>unclear</td>
<td>15</td>
</tr>
<tr>
<td>user-system interaction measurements</td>
<td>10</td>
</tr>
<tr>
<td>counting occurrences in text</td>
<td>8</td>
</tr>
<tr>
<td>user-text interaction measurements</td>
<td>6</td>
</tr>
<tr>
<td>other</td>
<td>1</td>
</tr>
</tbody>
</table>

_Howcroft et al., 2020_
How is it being measured?

Positive and Negative Framing
*How much more fluent is sentence A versus sentence B?*
→ implicitly prime rater that A is better than B

Demand Characteristics
*We consider sentences that end with “.” as more formal than sentences that end with “!”*
→ Biases raters to pay more attention to model artifacts

Anchoring and Adjusting
*Select sentences from model A as examples in the instruction*
→ Biases raters to prefer outputs that look like A over B.

<table>
<thead>
<tr>
<th>Form</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>direct quality estimation</td>
<td>207</td>
</tr>
<tr>
<td>relative quality estimation</td>
<td>72</td>
</tr>
<tr>
<td>(dis)agreement with quality statement</td>
<td>48</td>
</tr>
<tr>
<td>classification</td>
<td>38</td>
</tr>
<tr>
<td>task performance measurements</td>
<td>35</td>
</tr>
<tr>
<td>qualitative feedback</td>
<td>20</td>
</tr>
<tr>
<td>evaluation through post-editing/annotation</td>
<td>18</td>
</tr>
<tr>
<td>unclear</td>
<td>15</td>
</tr>
<tr>
<td>user-system interaction measurements</td>
<td>10</td>
</tr>
<tr>
<td>counting occurrences in text</td>
<td>8</td>
</tr>
<tr>
<td>user-text interaction measurements</td>
<td>6</td>
</tr>
<tr>
<td>other</td>
<td>1</td>
</tr>
</tbody>
</table>

*Howcroft et al., 2020*

Schoch et al., 2020
How many annotations do we need?

Humans measure the “true” difference between two systems, but have **high variance**. Metrics have lower variance, but are **biased**. Both are sources of errors.

As models get better, the differences between them get smaller. As a result, we need more annotator judgements.

To detect a difference of 1 point on a 1-100 scale in WMT, we need 10,000 perfect annotator judgements.

Yet, most annotations in my survey had n=100 or smaller.
Who is measuring?
And why may this be a problem?

Some aspects are easier to assess without professional raters (linguistic quality vs. content quality).

Crowdworkers tend to much have a higher variance than professional raters.

Agreement between ratings produced by linguists and those from crowdworkers can be extremely low.

Table 4: Linear regression is used to model Overall Quality scores as a function of judges, topics, and systems, respectively, for each data set. The $R^2$ values, which give the fraction of variance explained by each of the six models, are shown.

MTurk workers also had a much higher correlation between linguistic and overall quality than experts.  

Freitag et al., 2021  

Gillick + Liu, 2010
We need methods to deal with noisy ratings.

Text Analysis Conference Summarization track evaluation:
- Each assessor is assigned to a topic and evaluates all summaries, even duplicate ones
- We can identify within-annotator consistency

CLASSY is a (non-neural!) logistic regression model trained on these ratings

→ Excluding the most inconsistent annotated data can lead to higher correlation.

<table>
<thead>
<tr>
<th>Pyramid</th>
<th>NoModels</th>
<th></th>
<th>AllPeers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>main</td>
<td>update</td>
<td>main</td>
<td>update</td>
</tr>
<tr>
<td>CLASSY1_Pyr</td>
<td>0.956</td>
<td>0.898</td>
<td>0.945</td>
<td>0.936</td>
</tr>
<tr>
<td>CLASSY1_Pyr_new</td>
<td>0.950</td>
<td>0.895</td>
<td>0.932</td>
<td>0.955</td>
</tr>
<tr>
<td>CLASSY1_Pyr_new</td>
<td>0.960</td>
<td>0.900</td>
<td>0.940</td>
<td>0.955</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responsiveness</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NoModels</td>
<td></td>
<td>AllPeers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>main</td>
<td>update</td>
<td>main</td>
<td>update</td>
</tr>
<tr>
<td>CLASSY2_Resp</td>
<td>0.951</td>
<td>0.903</td>
<td>0.948</td>
<td>0.963</td>
</tr>
<tr>
<td>CLASSY2_Resp_new</td>
<td>0.954</td>
<td>0.907</td>
<td>0.973</td>
<td>0.950</td>
</tr>
<tr>
<td>CLASSY4_Resp</td>
<td>0.951</td>
<td>0.927</td>
<td>0.830</td>
<td>0.949</td>
</tr>
<tr>
<td>CLASSY4_Resp_new</td>
<td>0.943</td>
<td>0.928</td>
<td>0.887</td>
<td>0.946</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Readability</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NoModels</td>
<td></td>
<td>AllPeers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>main</td>
<td>update</td>
<td>main</td>
<td>update</td>
</tr>
<tr>
<td>CLASSY3_Read</td>
<td>0.768</td>
<td>0.705</td>
<td>0.844</td>
<td>0.907</td>
</tr>
<tr>
<td>CLASSY3_Read_new</td>
<td>0.793</td>
<td>0.721</td>
<td>0.858</td>
<td>0.906</td>
</tr>
</tbody>
</table>

Numbers are correlation between output and measure of the subsection.
What do noisy ratings mean for metrics?

**Surprise #1**
Metrics agree more with the high-quality annotations than with noisy ones, despite being trained on noisy annotations.

**Surprise #2**
Metrics have a higher agreement with MQM than WMT has with MQM.

**Surprise #3** not really
Previous findings about metric quality rankings are wrong.

---

Note that this is for MT, not summarization.
What can we do about this?

We don’t know what is being measured and how.
→ Write human evaluation datasheets \(\text{(Shimorina + Belz, 2021)}\)

None of our results are statistically significant.
→ Estimate the effect size before running evaluations and use significance tests to verify results.

Expert raters provide much better results than crowdworkers.
→ Verify crowdsourcing results multiple times and think what qualifications are required for what you want to measure.

→ Don’t treat human evaluation as the ultimate answer to all evaluation problems
So far...

- Our metrics don’t measure what we want (at least not well).
- Human evaluation can help evaluate models and develop metrics, but only in theory.

What about our datasets?

What does a better score on dataset X mean?
Datasets are broken
So far…

- Our metrics don’t measure what we want (at least not well).
- Human evaluation can help evaluate models and develop metrics, but only in theory.

What about our datasets?
What does a better score on dataset X mean?

And what is this dataset X?
Let’s look at the survey.
What does the survey tell us?

- 27 different datasets in 20 papers
- Only two non-English datasets
- CNN/DM remains the most popular dataset

→ How can we as a field make progress on improving summarization if we don’t have a (good) standard benchmark?
→ Also, how can we say we are making progress if we focus on a single language?
Problem #1: Noise

Reference summaries often contain extraneous information, such as hyperlinks and click-bait descriptions of other articles.

Raters prefer lead-3 over the CNN/DM reference.

→ Can we expect faithful models if our data is not?

read: falcao still ' has faith ' that he could continue at man utd next season. click here for the latest manchester united news.

Doesn’t that make the whole CNN/DM task pointless?

<table>
<thead>
<tr>
<th>Models</th>
<th>Hallucinated</th>
<th>Faith.</th>
<th>+Fact.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>E</td>
<td>I ∪ E</td>
</tr>
<tr>
<td>GOLD</td>
<td>7.4</td>
<td>73.1</td>
<td>76.9</td>
</tr>
</tbody>
</table>

Fabbri et al., 2021, Stiennon, Ouyang, Wu, Ziegler et al., 2020, Maynez, Narayan, et al., 2020
Problem #2: Splits

Results look completely different depending on how the test set was constructed.

A good model should do well on all expected data during deployment in a live scenario. Not just i.i.d. data.

<table>
<thead>
<tr>
<th>Task</th>
<th>Model</th>
<th>Standard</th>
<th>Random</th>
<th>Heuristic</th>
<th>Adversarial</th>
<th>New Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>HEADLINE GENERATION*</td>
<td>seq2seq</td>
<td>0.073</td>
<td>0.095</td>
<td>0.062</td>
<td>0.040</td>
<td>0.069</td>
</tr>
</tbody>
</table>

Søgaard et al., 2021, Ribeiro et al., 2020
Problem #2: Splits

Results look completely different depending on how the test set was constructed.

A good model should do well on all expected data during deployment in a live scenario. Not just i.i.d. data.

But, most datasets only have one test set. How do we test calibration?

→ We need focused challenge sets to test capabilities.

Søgaard et al., 2021, Ribeiro et al., 2020
Problem #3: New Concepts

Training sets usually remain static, but real test data does not.

How does a model perform for new concepts?

We created 3 test sets for pre-2020 datasets:
- XSum (En)
- MLSum (De)
- MLSum (Es)

Original collection method, but COVID-19 related articles.

Gehrmann et al., 2021, Mille et al., 2021
Problem #4: Style

Performance should not depend on the reference style.

We split the XSum test set into 10 buckets depending on reference abstractiveness.

The more abstractive a reference, the lower the score.

Similar finding in MT (Freitag et al., 2020)

Gehrmann et al., 2021, Mille et al., 2021
What can we do about this?

→ Document limitations, issues, and social impact (Gebru et al., 2018, Bender + Friedman, 2018).

→ Create evaluation suites instead of i.i.d. test sets.

→ Evaluate worst-case performance, not only average.

→ Think of a dataset, its splits, and documentation as a “living” object instead of a static entity.

¹We released an NLG-specific template in McMillan-Major et al., (2021)
04

So how do we fix things?
We need to break through this circular dependency.

At the moment, we can’t identify whether and how our models fail, or whether failure is attributable to the data, model, or evaluation.

→ A single researcher cannot solve every problem. We thus need easy-to-use infrastructure to stay up to date with the latest developments, combining everyone’s strengths.
This is what we are trying with the **Generation**, **Evaluation**, and **Metrics** Benchmark.

Instead of dictating what should be used, let’s make it easy to explore the right way to do things.
What does building infrastructure entail?

**Data**
- Select and document tasks
  - Build
- Correct and add splits
  - Improve
- Collect new datasets
  - Expand

**Automatic Metrics**
- Shared metrics environment
  - Build
- Fine-grained breakdowns
  - Improve
- Develop new metrics
  - Expand

**Human Evaluation**
- One framework for all tasks
  - Build
- Consistent definitions and quality control
  - Improve
- Build annotation corpus
  - Expand

**Models**
Let’s unbreak data

Develop transformations and filters of datasets to test robustness and performance on subpopulations.

Instead of chasing the highest number, try to break models.

More infos at https://gem-benchmark.com/nl_augmenter
Also, https://robustnessgym.com and many others.
Let's unbreak data

Develop transformations and filters of datasets to test robustness and performance on subpopulations.

Instead of chasing the highest number, try to break models.

More infos at [https://gem-benchmark.com/nl_augmenter](https://gem-benchmark.com/nl_augmenter) 🦜 → 🐍 Also, [https://robustnessgym.com](https://robustnessgym.com) and many others.

**LIST OF TASKS**

The list below links to data statements [1, 2] for each of the datasets that are part of GEM tasks. The template used to produce the statements and a guide on how to write them can be found here: [download template] [view guide].

- **MLSum** summarization
  Large-scale multilingual dataset for evaluating abstractive summarization.
- **XSum** Summarization
  Large scale monolingual dataset for evaluating extreme summarization.
- **WikiLingua** Summarization
  Large-scale multilingual dataset for evaluating cross-lingual abstractive summarization.
- **WebNLG** Structure-to-text
  The WebNLG dataset is a large bi-lingual dataset with crowdsourced reference texts and a rather large variety of knowledge in the inputs. A web-based evaluation platform is already existing.
- **CommonGen** Structure-to-text
  A medium sized corpus with a unique reasoning challenge and interesting evaluation possibilities.
- **E2E** Structure-to-Text
  One of the largest limited-domain NLG datasets and is frequently used as a data-to-text generation benchmark.
- **DART** Structure-to-Text
  Hierarchical, structured format with its open-domain nature.
- **Czech Restaurant** Structure-to-Text
  One of a few non-English data-to-text datasets in a well-known domain, covering a morphologically rich language.
- **ToTTo** Structure-To-Text
  Controlled Table2Text task with non-divergent, annotator-revised text outputs.
- **Wiki-Auto** Simplification
  Wiki-Auto is the largest open text simplification dataset currently available. For GEM, Wiki-Auto acts as the training set.
- **TURKCorpus** Simplification
  TURKCorpus is a high-quality simplification dataset where each source sentence is associated with 8 human-written simplifications.
- **ASSET** Simplification
  ASSET is a high quality simplification dataset where each source (not simple) sentence is associated with 10 human-written simplifications.
- **Schema-Guided Dialog** dialog
  Modeling task-oriented dialog.

Currently 13 documented tasks in 18 languages. Support for loaders in 😊 Datasets and TFDS. Soon 30+ tasks across 40+ languages.

More at [https://gem-benchmark.com/data_cards](https://gem-benchmark.com/data_cards)
Let’s unbreak metrics

We can use multiple metrics instead of only ROUGE.

Our library computes 100+ statistics and metrics for any generation task. For supported tasks, it provides fine-grained breakdowns.

The library has support for caching, runs non-GPU metrics in parallel, and we are adding many more metrics.

We are hoping that it will make the lives of model developers and metrics researchers easier.

Putting this together - we can develop performance and robustness reports.

**Question**: How robust is my model to punctuation mistakes?

**Answer**: By framing robustness in causal terms and measure multiple response metrics, we can audit models without perfect metric.
Final Lessons

We don’t really know how to evaluate models...

But we can do a better job at evaluation

- We can write better documentation
- We can report more metrics
- We can frame model results around their robustness

Instead of aiming for higher ROUGE numbers, let’s audit models, evaluation approaches, and datasets.

Sebastian Gehrmann
gehrmann@google.com
@SebGehr
Google Research