NLG Evaluation

Ehud Reiter, Univ of Aberdeen
Contents

▪ Concepts
  ▪ Extrinsic evaluation
  ▪ Intrinsic evaluation
  ▪ Metric evaluation
  ▪ Evaluating hallucination
  ▪ Evaluating explanations
What is Evaluation?

- Experimentally testing hypotheses
  - Is system/algorithm/model/etc X better than baseline or state-of-the-art?
    - NLG: texts more useful, easier to read, etc
  - Is system/algorithm/model/etc. X useful in real-world applications?
- Scientific rigour is essential!
Types of NLG Evaluation

- **Extrinsic**: Ask people to use a system, see if it helps them
- **(Human) intrinsic**: Ask people to rate and assess NLG texts
- **Metric**: Compare NLG texts to human-written “reference” texts
Quality criteria

- **What evaluated**: correctness, goodness, specific features/aspects
- **Aspect evaluated**: form (eg grammar), content (eg, meaning), both
- **Frame of reference**: text on own, relative to input, in external task context
Replicability

- Can evaluations be repeated by other researchers?
  - Not good science if not replicable!
- Just starting a research project on this, let me know if interested!
  - Looking for partners to repeat experiments
Another perspective

Good general paper/talk on NLG eval


- Video:  
  https://www.youtube.com/watch?v=eSuefO4CkHQ
Contents

- Concepts
- Extrinsic evaluation
- Intrinsic evaluation
- Metric evaluation
- Evaluating hallucination
- Evaluating explanations
Extrinsic Evaluation

- Directly measure impact of NLG system on task performance or some other “extrinsic” outcome measure.
- Does system achieve its communicative goal?
Example: Summarise consult

- Goal: Generate a summary of a doctor-patient consultation, for the medical record.
  - Summary is post-edited by doctor before being entered into record
### Example

<table>
<thead>
<tr>
<th>Transcript</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinician</td>
<td>Hello.</td>
</tr>
<tr>
<td>Patient</td>
<td>Hello, how are you?</td>
</tr>
<tr>
<td>Clinician</td>
<td>Hello. How can I help you this morning?</td>
</tr>
<tr>
<td>Patient</td>
<td>All right. I just had some diarrhea for the last three days and it’s been affecting me. I need to stay close to the toilet. And yeah, it’s been affecting my day-to-day activities.</td>
</tr>
<tr>
<td>Clinician</td>
<td>I’m sorry to hear that and when you say diarrhea, what do you mean by diarrhea? Do you mean you’re going to the toilet more often or are your stools more loose?</td>
</tr>
<tr>
<td>Patient</td>
<td>Yeah, so it’s like loose and watery <strong>stole</strong> going to the toilet quite often.</td>
</tr>
<tr>
<td>Clinician</td>
<td><strong>freak</strong> ...</td>
</tr>
<tr>
<td></td>
<td>3/7 hx of diarrhea, mainly watery.</td>
</tr>
<tr>
<td></td>
<td>No blood in stool. Opening bowels x6/day.</td>
</tr>
<tr>
<td></td>
<td>Associated LLQ pain - crampy, intermittent, nil radiation.</td>
</tr>
<tr>
<td></td>
<td>Also vomiting - mainly bilous.</td>
</tr>
<tr>
<td></td>
<td>No blood in vomit.</td>
</tr>
<tr>
<td></td>
<td>Fever on first day, nil since.</td>
</tr>
<tr>
<td></td>
<td>Has been feeling lethargic and weak since.</td>
</tr>
<tr>
<td></td>
<td>Takeaway 4/7 ago - Chinese restaurant.</td>
</tr>
<tr>
<td></td>
<td>Wife and children also unwell with vomiting, but no diarrhea. No other unwell contacts.</td>
</tr>
<tr>
<td></td>
<td>PMH: Asthma</td>
</tr>
<tr>
<td></td>
<td>DH: Inhalers</td>
</tr>
<tr>
<td></td>
<td>SH: works as an accountant.</td>
</tr>
<tr>
<td></td>
<td>Lives with wife and children.</td>
</tr>
<tr>
<td></td>
<td>Affecting his ADLs as has to be near toilet.</td>
</tr>
<tr>
<td></td>
<td>Nil smoking/etOH hx</td>
</tr>
</tbody>
</table>
Evaluation

- Ask doctors to listen to a consultation, view draft summary, and post-edit to fix mistakes
  - Time post-editing

- Also categorise mistakes in draft summary
  - Incorrect statements vs omission
  - Critical vs non-critical
Evaluation Process

Dataset

Case card read by Actor (Patient) joins Medical consultation carries out Consulting Clinician writes Note human_note

Speech-to-text Transcript

audio

Human Evaluation

1) Listen to consultation and write own note writes Note eval_note

2) Post-edit 4 generated + 1 human note produces Generated Notes

3) List incorrect statements and omissions

4) Qualitative feedback

Note edited_note
Evaluation criteria

- Post-edit time: 136s
  - Big chunk of 10-min consultation!
  - But still faster than manual writing

- Errors
  - 3.9 incorrect statements per summary (average)
  - 6.6 omissions per summary (average)
  - Post-editing takes time because of num of errors.
## Error Types

<table>
<thead>
<tr>
<th>Issue</th>
<th>Examples</th>
<th>Occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discourse level</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contradiction</td>
<td><em>no family history of bowel issues.</em> <em>father has history of colon cancer</em></td>
<td>25</td>
</tr>
<tr>
<td>Contradiction not reported</td>
<td>patient corrected herself but note did not pick it up.</td>
<td>4</td>
</tr>
<tr>
<td>Symptom mentioned is reported as fact</td>
<td><em>tingling of hands</em> stated by clinician (not refuted/confirmed by pt)</td>
<td>18</td>
</tr>
<tr>
<td>Misleading statement</td>
<td>statement: <em>not working</em> when patient has been off ill for a few days due to current sickness reads like patient is unemployed</td>
<td>9</td>
</tr>
<tr>
<td><strong>Factual errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hallucination</td>
<td><em>at home and in a private place</em> was not mentioned in the consultation</td>
<td>17</td>
</tr>
<tr>
<td>Incorrect statement</td>
<td>statement: <em>brother has diabetes.</em> Correction: <em>mother has diabetes</em></td>
<td>34</td>
</tr>
<tr>
<td>Nonsensical</td>
<td><em>No recent unwell with diarrhoea</em></td>
<td>18</td>
</tr>
<tr>
<td><strong>Stylistic errors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repetition</td>
<td><em>loose and watery stools</em> <em>stool is mainly watery</em></td>
<td>93</td>
</tr>
<tr>
<td>Incorrect order of statements</td>
<td><em>heart attack</em> should be in PMH; structure of history a bit jumbled; recorded social smoker in alcohol section.</td>
<td>38</td>
</tr>
<tr>
<td>Use of not universally recognised acronyms</td>
<td><em>NRS/EMS/DOA are not standard acronyms</em></td>
<td>7</td>
</tr>
<tr>
<td><strong>Omissions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Generic</td>
<td>no mention of <em>unable to open bowels</em></td>
<td>57</td>
</tr>
<tr>
<td>Omissions of important negatives</td>
<td><em>No fever</em> <em>No shortness of breath</em></td>
<td>5</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good behaviour</td>
<td>Contains all the history that was covered in the audio and follows a logical structure</td>
<td>21</td>
</tr>
</tbody>
</table>
Outcome

- A lot of errors, of many different types
  » Neural NLG systems make mistakes!
- Research question:
  » Which model better: focus on post-edit time
  » Is system useful: Need to embed in workflow, UI/UX which allows easy/fast post-editing
Other examples

- Decision support: does an NLG system help doctors make better decisions?

- Behaviour change: does an NLG system help people to stop smoking?
Contents

- Concepts
- Extrinsic evaluation
- Intrinsic evaluation
- Metric evaluation
- Evaluating hallucination
- Evaluating explanations
(Human) Intrinsic Evaluation

- Ask human evaluators to rate texts on various quality criteria
  - User opinion, not based on task performance
Example: Babytalk Nurse

- BT-Nurse: Generated shift summaries for nurses starting shift in neonatal ICU
  - Evaluated by asking nurses what they thought of the texts
Context: Neonatal ICU
Input: Sensor Data
BT-Nurse: Example

- Extract from 5 page report

  Respiratory Support

  Current Status

  ...

  SaO2 is variable within the acceptable range and there have been some desaturations.

  ...

  Events During the Shift

  A blood gas was taken at around 19:45. Parameters were acceptable. pH was 7.18. CO2 was 7.71 kPa. BE was -4.8 mmol/L.
BT-Nurse: Evaluation

- Hypothesis: Nurses will find BT-Nurse texts to be understandable, accurate, and helpful
  » Not measuring medical outcome
- Subjects: Neonatal ICU nurses
  » 165 trials, where a nurse read a BT-Nurse texts
  » 54 nurses, most participated in multiple trials
- Material: BT-Nurse texts
  » No control/baseline
Experimental Design

- Procedure (for incoming nurses)
  » Research nurse vetted BT-Nurse text, to screen out texts which could harm patient care (ethics)
    - In fact no BT-Nurse texts were screened out
  » Duty nurse read BT-Nurse text
  » Nurse rated texts understandable, accurate, helpful (3-pt)

- Analysis
  » Percentage of nurses rated texts understand, etc
  » Quantitative summary of free-text comments
Results

- Numerical results
  - 90% of texts understandable
  - 70% of texts accurate
  - 60% of texts helpful
  - [no texts rejected as potentially harmful]
  - All numbers are statistically significant

- Many free-text comments
  - Most common was request for more information
  - A few “really helped me” comments
  - Some comments highlighted software bugs
Good exper design!

- Good experimental design matters!
  » Subjects, material, questions, stats, etc
- Poorly designed human eval worthless
Contents

- Concepts
- Extrinsic evaluation
- Intrinsic evaluation
- **Metric evaluation**
- Evaluating hallucination
- Evaluating explanations
Metric evaluation

- Evaluate NLG system automatically, usually by comparing output texts to human-written “reference” texts
- Various algorithms for comparison: BLEU, ROUGE, BERTSCORE, etc
- Validity: Metrics only useful if they reliably predict (high-quality) human evaluations
Example: Weather

- Task: NLG weather forecasts for oil rigs
- Evaluation
  - Human intrinsic: Ask workers in industry to evaluate NLG weather forecasts
  - Metric: used popular metrics to eval NLG weather forecasts
  - Research question: do metrics predict human evaluation?
Metric example

- Human
  - SSW 16-20 GRADUALLY BACKING SSE THEN BECOMING VARIABLE 10 OR LESS BY MIDNIGHT

- SumTime (NLG)
  - SSW’LY 16-20 GRADUALLY BACKING SSE’LY THEN DECREASING VARIABLE 4-8 BY LATE EVENING
  - SSW’LY 16-20 GRADUALLY BACKING SSE’LY THEN DECREASING BECOMING VARIABLE 4-8 10 OR LESS BY LATE EVENING MIDNIGHT

- Compute score using BLEU, edit distance, etc
Experimental Design

- **Procedure**
  - Use BLEU, ROUGE, etc metrics to measure similarity of NLG texts to human reference forecasts
    - Several NLG systems
  - Also do human intrinsic eval on these sys

- **Analysis**
  - Do metrics predict (correlate with) human evaluations?
## Results

Table 5  
Experiment 1: Metric scores against three reference texts (produced by rewriting corpus texts), for the set of 18 forecasts used in expert evaluation.

<table>
<thead>
<tr>
<th>System</th>
<th>Exp</th>
<th>Non</th>
<th>NIST-5</th>
<th>BLEU-4</th>
<th>ROUGE-SU4</th>
<th>ROUGE-2</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST-Hybrid</td>
<td>3.82</td>
<td>3.90 (1)</td>
<td>6.382 (3)</td>
<td>0.584 (4)</td>
<td>0.558 (4)</td>
<td>0.528 (4)</td>
<td>0.705 (5)</td>
</tr>
<tr>
<td>pCRU-greedy</td>
<td>3.59</td>
<td>3.51 (3)</td>
<td>6.871 (2)</td>
<td>0.694 (2)</td>
<td>0.656 (2)</td>
<td>0.634 (2)</td>
<td>0.800 (2)</td>
</tr>
<tr>
<td>ST-Corpus</td>
<td>3.22</td>
<td>3.62 (2)</td>
<td>8.705 (1)</td>
<td>0.951 (1)</td>
<td>0.839 (1)</td>
<td>0.815 (1)</td>
<td>0.917 (1)</td>
</tr>
<tr>
<td>pCRU-roulette</td>
<td>3.11</td>
<td>3.49 (4)</td>
<td>6.206 (4)</td>
<td>0.563 (5)</td>
<td>0.554 (5)</td>
<td>0.51 (5)</td>
<td>0.735 (4)</td>
</tr>
<tr>
<td>pCRU-2gram</td>
<td>2.68</td>
<td>3.29 (5)</td>
<td>5.925 (5)</td>
<td>0.598 (3)</td>
<td>0.586 (3)</td>
<td>0.556 (3)</td>
<td>0.783 (3)</td>
</tr>
<tr>
<td>pCRU-random</td>
<td>2.43</td>
<td>2.51 (6)</td>
<td>4.608 (6)</td>
<td>0.355 (6)</td>
<td>0.462 (6)</td>
<td>0.419 (6)</td>
<td>0.649 (6)</td>
</tr>
</tbody>
</table>

For convenience, expert (Exp) and non-expert (Non) scores are also shown.
Results

- Metrics did not predict human eval
  - System preferred by humans (SumTime-Hybrid) was rated fourth by BLEU, etc
Not unusual!

- Lots of papers find poor correlation of metrics with human eval!
  - Gehrmann et al (2022)
  - Moramarco et al (2022)
Lesson

- Metrics do NOT predict (or correlate with) human eval of NLG systems!
- Should never be primary evaluation of NLG systems
  » Can include as secondary eval
Contents

▪ Concepts
▪ Extrinsic evaluation
▪ Intrinsic evaluation
▪ Metric evaluation
▪ Evaluating hallucination
▪ Evaluating explanations
Evaluating Accuracy

- Accuracy (hallucination) is big problem
  » Especially in neural NLG
  » Especially in longer texts
- Users expect NLG texts to be accurate!
  » Lose trust if sys produces inaccurate texts
- How do we evaluate accuracy?
Example: basketball stories

- Accuracy of summaries of basketball games
  - Produced from “box score” game data
  - 300 words on average

# Team & Player Data

<table>
<thead>
<tr>
<th>TEAM</th>
<th>W</th>
<th>L</th>
<th>H1-PTS</th>
<th>H2-PTS</th>
<th>PTS</th>
<th>FG%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grizzlies</td>
<td>5</td>
<td>0</td>
<td>46</td>
<td>56</td>
<td>102</td>
<td>.486</td>
</tr>
<tr>
<td>Suns</td>
<td>3</td>
<td>2</td>
<td>52</td>
<td>39</td>
<td>91</td>
<td>.559</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Player</th>
<th>TEAM</th>
<th>PTS</th>
<th>REB</th>
<th>AST</th>
<th>BLK</th>
<th>STL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marc Gasol</td>
<td>Grizzlies</td>
<td>18</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Isaiah Thomas</td>
<td>Suns</td>
<td>15</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
The Memphis Grizzlies (5-2) defeated the Phoenix Suns (3-2) Monday 102-91 at the Talking Stick Resort Arena in Phoenix. The Grizzlies had a strong first half where they out-scored the Suns 59-42. Marc Gasol scored 18 points, leading the Grizzlies. Isaiah Thomas added 15 points, he is averaging 19 points on the season so far.
The Memphis Grizzlies (5-2) defeated the Phoenix Suns (3-2) **Monday** 102-91 at the **Talking Stick Resort Arena** in Phoenix. The Grizzlies had a **strong** first half where they **out-scored** the Suns 59-42. Marc Gasol scored 18 points, **leading** the Grizzlies. **Isaiah Thomas** added 15 points, he is averaging **19** points on the season so far.
Mistake categories

<table>
<thead>
<tr>
<th>Name</th>
<th>Player, Team, day of week, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>Number, in any form.</td>
</tr>
<tr>
<td>Word</td>
<td>Word or phrase that is not Name/Number.</td>
</tr>
<tr>
<td>Context</td>
<td>Something that is contextually wrong.</td>
</tr>
<tr>
<td>Not Checkable</td>
<td>Impossible/time-consuming to check.</td>
</tr>
<tr>
<td>Other</td>
<td>Any other error.</td>
</tr>
</tbody>
</table>
Find mistake: Gold standard

- High-quality human eval to find mistakes
- Procedure
  - 3 (selected/vetted) Turkers annotate each text
  - Researcher combines (majority opinion)
- Process worked
  - High interannotator agreement
  - Various checks, including with domain experts
- Expensive
  - US$30 for each 300-word summary
Metrics

- Kasner et al (2021) proposed metric
  - Generate synthetic data with rule-based NLG
  - Train language model to detect errors (using real and synthetic data)
  - Best metric for acc detection in this domain

- Works well for simpler errors
- Not great for complex errors
## Kasner et al metric

<table>
<thead>
<tr>
<th>Type</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>0.75</td>
<td>0.85</td>
</tr>
<tr>
<td>Number</td>
<td>0.78</td>
<td>0.75</td>
</tr>
<tr>
<td>Word</td>
<td>0.51</td>
<td>0.48</td>
</tr>
<tr>
<td>Context</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>Not checkable</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>--</td>
</tr>
<tr>
<td>Overall</td>
<td>0.69</td>
<td>0.76</td>
</tr>
</tbody>
</table>
Contents

- Concepts
- Extrinsic evaluation
- Intrinsic evaluation
- Metric evaluation
- Evaluating hallucination
- Evaluating explanations
Evaluating explanations

- What are we interested in?
  - *Debugging*: Help ML engineers fix models
  - *Scrutability*: Help user find mistakes in AI reasoning
  - *Trust*: Increase user trust in system
  - Etc

How measure?

- What experimental designs are best for measuring scrutability, trust, etc?
- Can we use metrics to assess above?
- Poorly understood
Final Comments

- Be clear on what you are evaluating!
- Don’t use BLEU, ROUGE, etc!
- Extrinsic eval are best, but difficult
- Human intrinsic eval is often a sensible approach; must be well-designed!
- Many research challenges, especially in evaluating explanations