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
- Belz et al (2020). Disentangling the Properties of Human Evaluation Methods. *Proc of INLG-2020*

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

- Gehrman et al (2022). Repairing the Cracked Foundation. https://arxiv.org/abs/2202.06935
- Video: https://www.youtube.com/watch?v=eSuefO4CkHQ

Contents

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

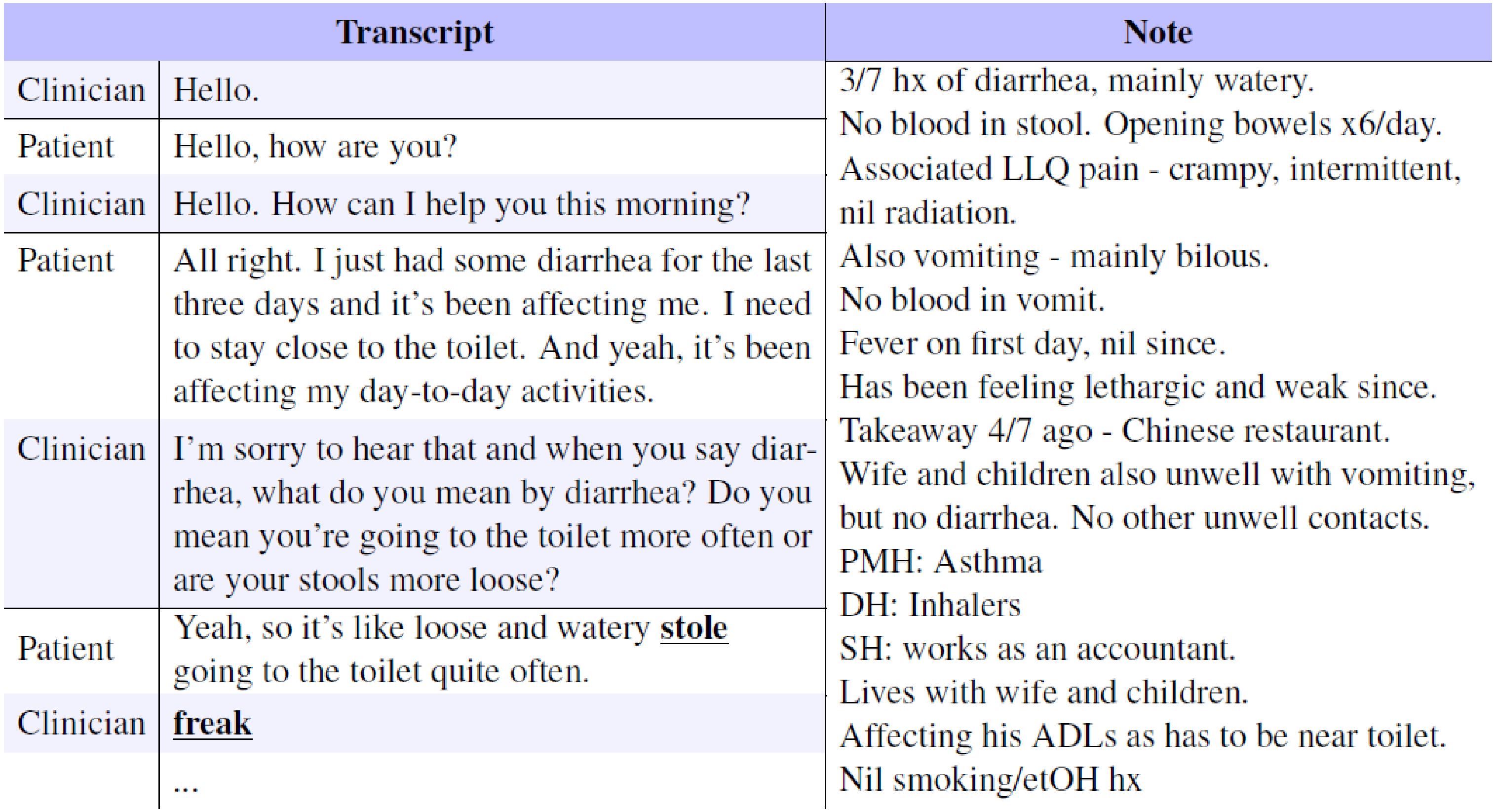
Extrinsic Evaluation

- 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
- F Moramarco et al (2022). Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. To appear in *Proc of ACL-2022.* (https://arxiv.org/abs/2204.00447)

Example

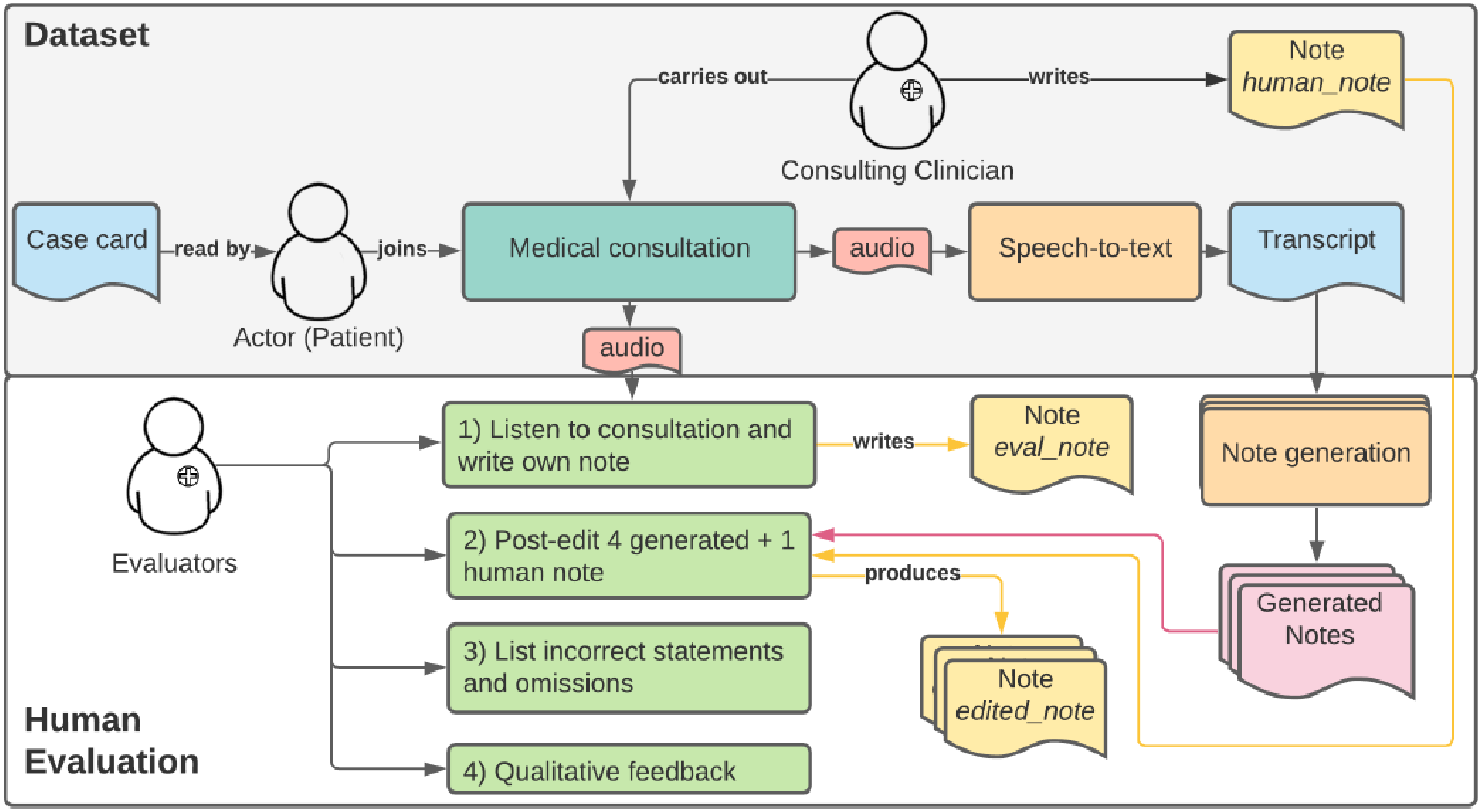


Transcript		Note
Clinician	Hello.	3/7 hx of diarrhea, mainly watery. No blood in stool. Opening bowels x6/day. Associated LLQ pain - crampy, intermittent, nil radiation. Also vomiting - mainly bilous. No blood in vomit. Fever on first day, nil since. Has been feeling lethargic and weak since. Takeaway 4/7 ago - Chinese restaurant. Wife and children also unwell with vomiting, but no diarrhea. No other unwell contacts. PMH: Asthma DH: Inhalers SH: works as an accountant. Lives with wife and children. Affecting his ADLs as has to be near toilet. Nil smoking/etOH hx
Patient	Hello, how are you?	
Clinician	Hello. How can I help you this morning?	
Patient	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.	
Clinician	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?	
Patient	Yeah, so it's like loose and watery **stole** going to the toilet quite often.	
Clinician	**freak**	
	...	

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 sum
  - » Incorrect statements vs omission
  - » Critical vs non-critical

Evaluation Process



Dataset

Consulting Clinician

carries out

writes

Note
human_note

Case card

read by

Actor (Patient)

joins

Medical consultation

audio

Speech-to-text

Transcript

audio

Human Evaluation

Evaluators

1) Listen to consultation and write own note

writes

Note
eval_note

Note generation

2) Post-edit 4 generated + 1 human note

produces

Generated Notes

3) List incorrect statements and omissions

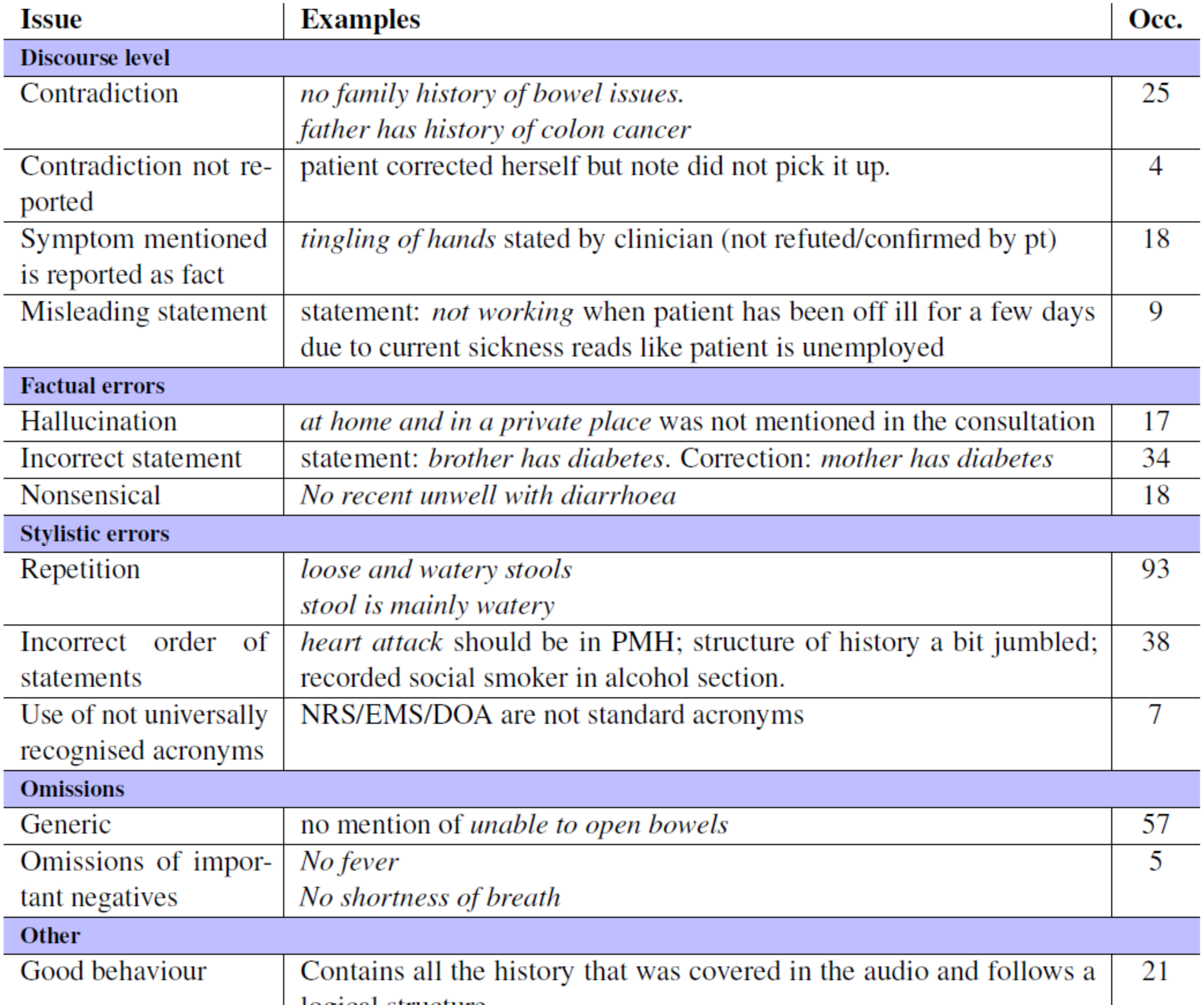
Note
edited_note

4) Qualitative feedback

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



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

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?
  - » F Portet et al (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence*
- Behaviour change: does an NLG system help people to stop smoking?
  - » E Reiter et al (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence*

Contents

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

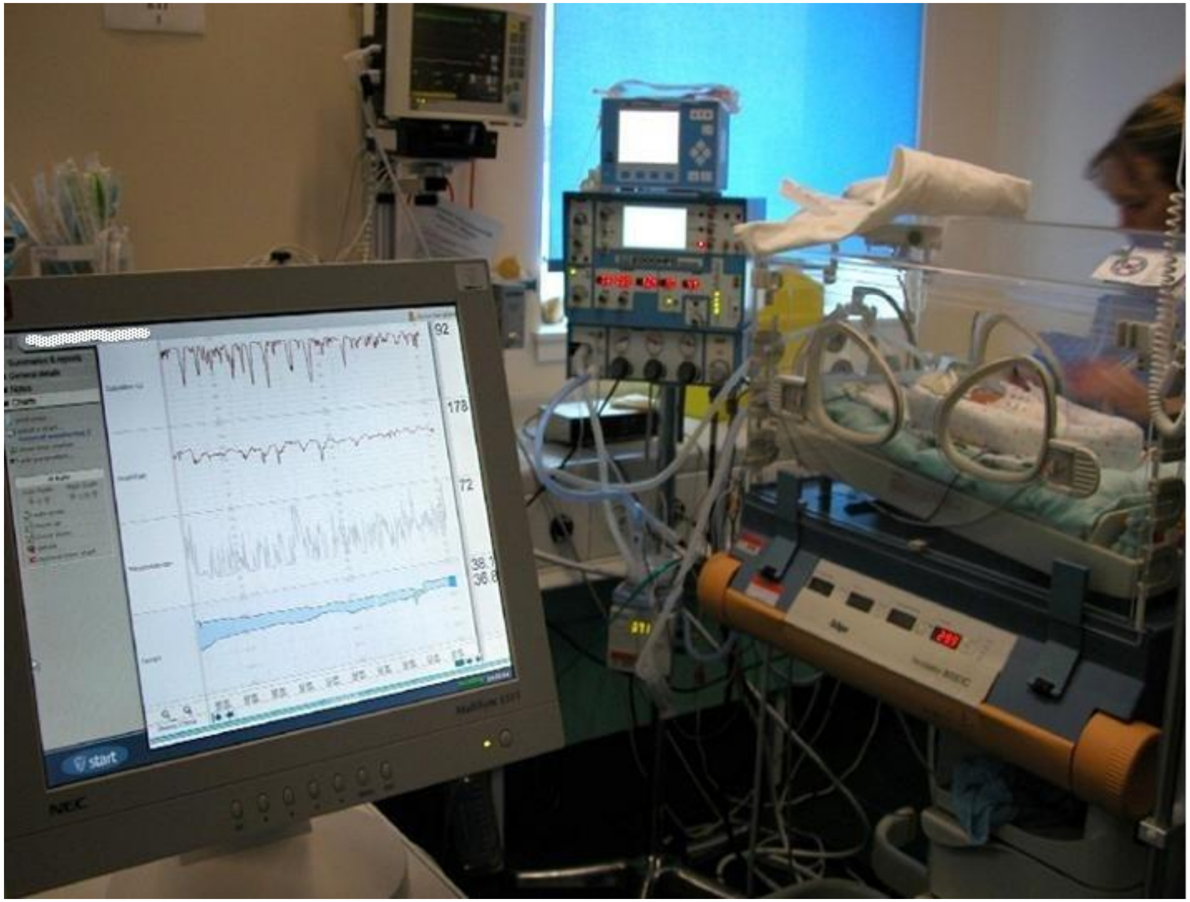
(Human) Intrinsic Evaluation

- Ask human evaluators to rate texts on various quality criteria
  - » User opinion, not based on task performance

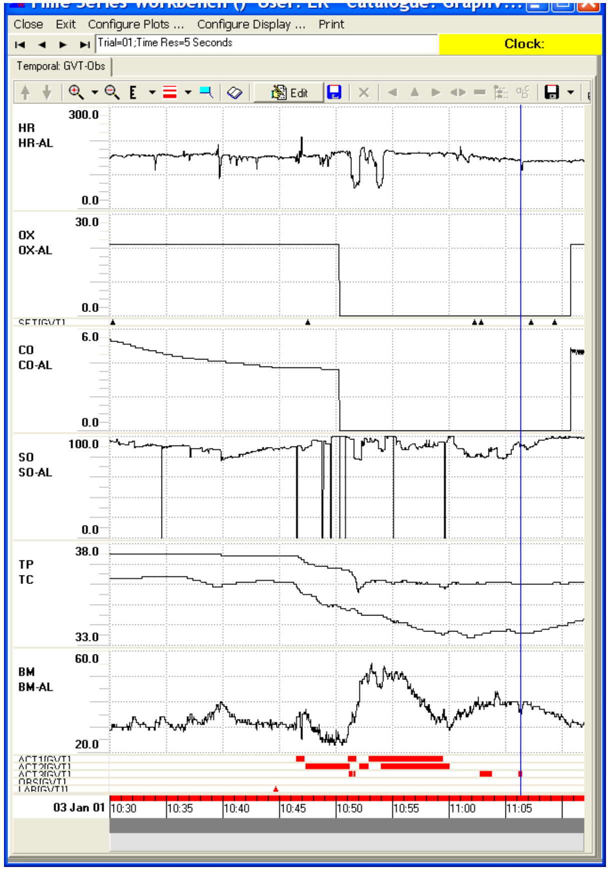
Example: Babytalk Nurse

- BT-Nurse: Generated shift summaries for nurses starting shift in neonatal ICU
  - » Evaluated by asking nurses what they thought of the texts
- J Hunter et al (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine*

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
- van der Lee et al (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer, Speech and Language*

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

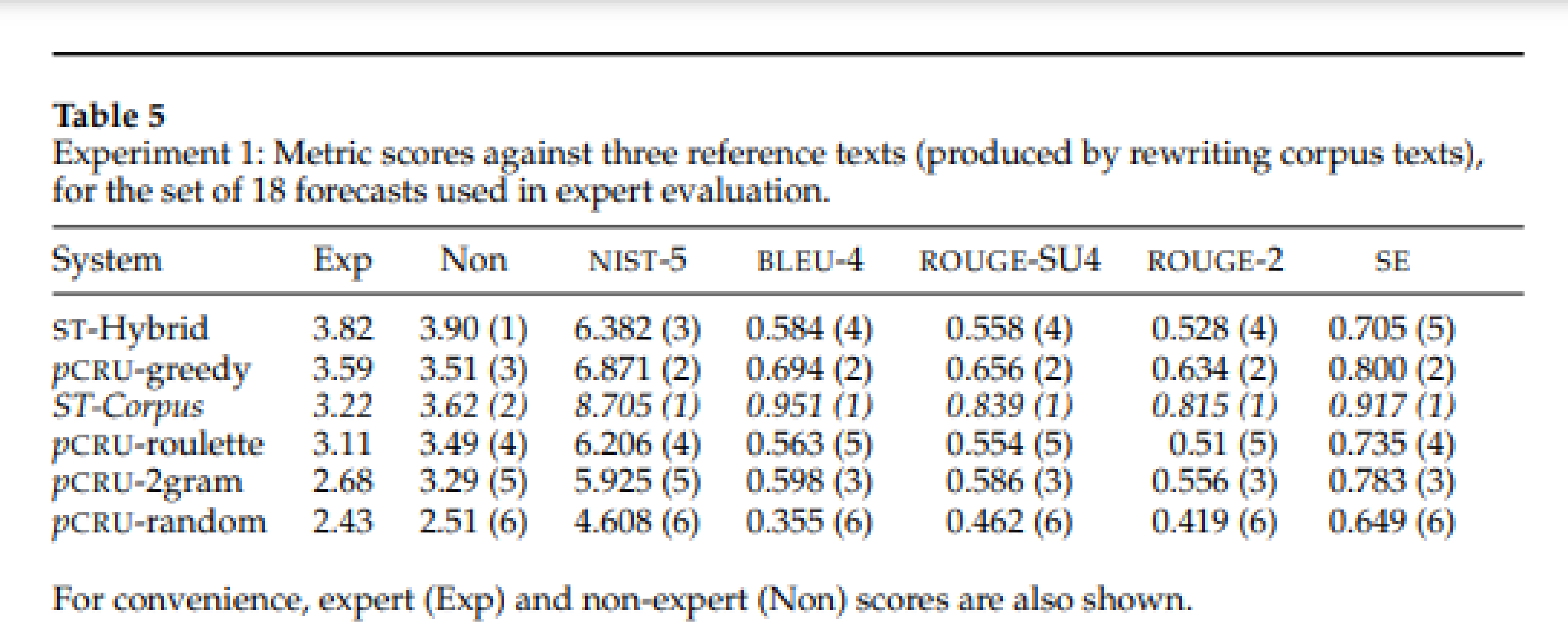


Table 5
Experiment 1: Metric scores against three reference texts (produced by rewriting corpus texts), for the set of 18 forecasts used in expert evaluation.

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

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
- E Reiter and A Belz (2009). An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*. https://aclanthology.org/J09-4008/

Not unusual!

- Lots of papers find poor correlation of metrics with human eval!
  - » Gehrmann et al (2022)
  - » Moramarco et al (2022)
  - » E Reiter (2018). A Structured Review of the Validity of BLEU. *Computational Linguistics*

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
- C Thomson, E Reiter (2020). A Gold Standard Methodology for Evaluating Accuracy in Data-To-Text Systems. *Proceedings of INLG-2020*.

Team & Player Data

TEAM	W	L	H1-PTS	H2-PTS	PTS	FG%
Grizzlies	5	0	46	56	102	.486
Suns	3	2	52	39	91	.559

Player	TEAM	PTS	REB	AST	BLK	STL
Marc Gasol	Grizzlies	18	5	6	0	4
Isaiah Thomas	Suns	15	1	2	0	1

NLG output (partial)

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.

Partial summary with errors

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

Name	Player, Team, day of week, etc.
Number	Number, in any form.
Word	Word or phrase that is not Name/Number.
Context	Something that is contextually wrong.
Not Checkable	Impossible/time-consuming to check.
Other	Any other error.

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
  - » Z Kasner et al (2021). Text-in-Context: Token-Level Error Detection for Table-to-Text Generation. *Proc of INLG-2021*
- Works well for simpler errors
- Not great for complex errors

Kasner et al metric

Type	Recall	Precision
Name	0.75	0.85
Number	0.78	0.75
Word	0.51	0.48
Context	0	--
Not checkable	0	--
Other	0	--
Overall	0.69	0.76

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
- N Tintarev and J Masthoff (2007). A survey of explanations in recommender systems. https://ieeexplore.ieee.org/iel5/4400942/4400943/04401070.pdf

How measure?

- What experimental designs are best for measuring scrutability, trust, etc?
- Can we use metrics to assess above?
- Poorly understood
- J Zhou et al (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. https://www.mdpi.com/2079-9292/10/5/593/pdf

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