High Quality Human Evaluation of NLG

Ehud Reiter

University of Aberdeen

Contents

- *High quality human evaluation*
- Old work
- Detecting accuracy errors
- Evaluating real-world utility of summaries
- Enhancing replicability
- Final thoughts

Evaluation in Medicine

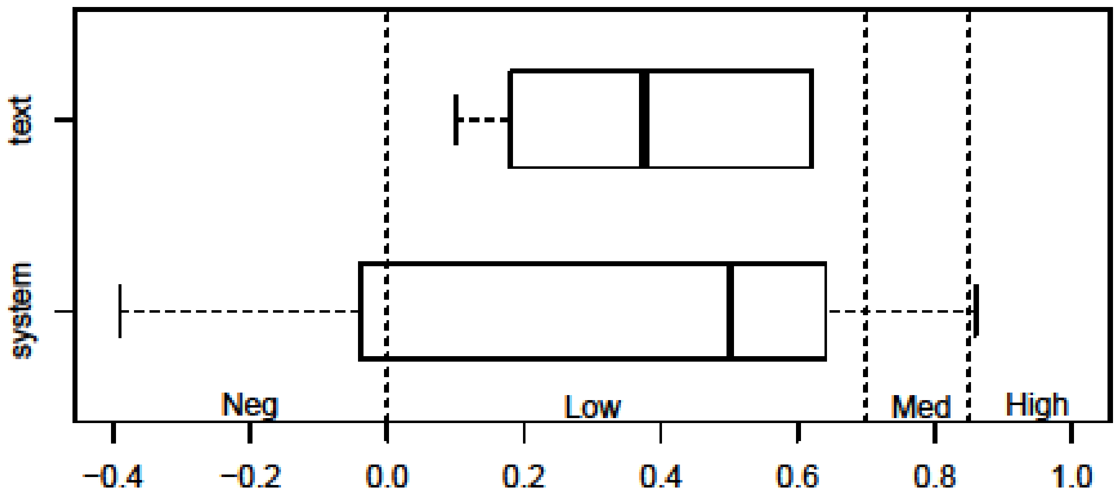
- Focus on high-quality expensive evaluations of clinical outcomes (RCTs)
- Sometimes can use cheaper/quicker *surrogate endpoints* for clinical outcome
  - » Eg, viral load instead of mortality
  - » Much quicker/easier to measure
  - » Only use if high correlation with clinical outcome
  - » Best studies avoid surrogate, use clinical outcomes

Evaluation in NLP

- Dominated by metrics (BLEU, etc)
  - » Metrics are surrogate endpoints
  - » Used even if limited corr with human eval
  - » Used everywhere, including top studies
- Human evaluations often limited
  - » Random crowdworkers as subjects
  - » Measure opinion rather than task outcome
- Need more high-quality human evals
  - » Analogous to RCT in medicine?

BLEU-human corr in NLG

- Meta-analysis across papers in ACL Anthology (Reiter 2018)



Human eval: subjects

- Most human evaluations in NLP use crowdworkers (eg Mechanical Turk)
- Freitag et al (2021): WMT human evals (based on monolingual crowdworkers) do NOT correlate well with structured evaluations by professional translators.

Human eval: opinion vs outcome

- Most human evaluations in NLG solicit ratings or opinions
- Usually what we really care about is whether NLP system helps people
  - » Task outcome (extrinsic eval)
- Rating/opinions may NOT correlate with task effectiveness
  - » Eg Law et al (2005)

Vision: High Qual Human Eval

- Do high-quality human eval of NLP
  - » Subjects with domain knowledge
  - » Objective/task outcome instead of opinion
- Use these for key experiments
- Use these to ground/validate metrics and cheaper human evals

Contents

- High quality human evaluation
- *Old work*
- Detecting accuracy errors
- Evaluating real-world utility of summaries
- Enhancing replicability
- Final thoughts

Smoking cessation

- NLG system generated stop-smoking leaflets based on user questionnaire
- Evaluated in medical-grade RCT
  - » 2500 subjects!
- Result: Simple fixed letter as effective as NLG letters
- Reiter et al (2003)

Clinical Decision Support

- NLG system summarized patient data for babies in neonatal ICU, to help clinicians decide on interventions
- Evaluation
  - » show clinicians NLG sum and visualisations
  - » asking them to make treatment decisions
  - » Compare decisions against gold stand
- Result: small diff, not stat significant
- Portet et al (2009)

Nursing Shift Handover

- NLG system generated nurse shift handover rep, for NICU babies
- Eval:
  - » System deployed, used on ward
  - » Researcher vets texts for errors
  - » Nurses say whether test useful
- Result: No serious errors, useful
- Hunter et al (2012)

Contents

- High quality human evaluation
- Old work
- *Example: Detecting accuracy errors*
- Evaluating real-world utility of summaries
- Enhancing replicability
- Final thoughts

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?
- Part of Craig Thomson’s PhD

Craig's work

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

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

Partial game summary

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.

Gold standard protocol

- High-quality human eval to find mistakes
  - » Thomson and Reiter (2020)
- Subjects
  - » Selected Mechanical Turk workers
  - » Know basketball, do well on vetting task
- Task
  - » Find and categorise mistakes
  - » More objective than 1-5 accuracy rating

Gold standard protocol

- Procedure
  - » 3 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

Cheaper Eval: Shared Task

- Created shared task to find cheaper and quicker techniques
  - » Should correlate with gold standard
- Cheaper human eval
- Metrics
- Thomson and Reiter (2021)

Quicker Human Eval

- Garneau and Lamontagne (2021): quicker and cheaper human eval
  - » Used metric to pre-annotate simple mistakes (not complex ones)
  - » Significant reduction in time/cost
  - » High agreement with gold stand
    - – Recall of .84
    - – Precision of .88

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)
- 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

Summary

- Identify area where good eval needed
  - » Evaluating accuracy is very important
- Created gold-standard human eval
  - » US$30 per text (expensive)
- Used gold standard to development metrics and cheaper human eval

Contents

- High quality human evaluation
- Old work
- Detecting accuracy errors
- *Evaluating real-world utility of summaries*
- Enhancing replicability
- Final thoughtrs

Evaluating Utility

- How evaluate if generated texts help users do tasks better or more quickly?
  - » Depends on task (and user)
- Part of Francesco Moramarco's PhD
  - » Task: summarizing patient-doctor consultations
  - » working with Babylon Health

Use case

- GP (doctor) talks to patient 5-10 mins
  - » Called “consultation”
- Needs to write summary of consultation
  - » For medical records, patient can see
- Currently done by GP
- Goal: NLP system gen draft summary
- Doctor “post-edits” to fix mistakes

Example

Consultation

Doctor: Hello? Good morning, Tim. Um, how can I help you this morning?

Patient: Um, so I’m having some, some pain, uh, in my tummy, like the lower part of my tummy. Um and I’ve just been feeling, quite, hot and sweaty.

Doctor: OK. Right, I’m sorry to hear that. When, when did your symptoms all start?

Patient: About two days ago.

Summary

Two days of lower abdominal pain.

How measure usefulness?

- Time spent post-editing NLP summary?
  - » Compared to time to write from scratch
- Quality of post-edited summary?
  - » determined by experienced clinician
- Number of mistakes in NLP summary
- Doctor satisfaction?
- Impact on workflow?

Not just averages

- Differences between doctors
  - » Post-editing time (and what is edited)
  - » Satisfaction
- Worst-case as well as average case
  - » No tolerance for medically misleading summaries

High Quality Human Eval

- Developing protocol
- Current version
  - » Doctors write their own summary
  - » Doctors shown NLP summary
  - » Doctors post-edit NLP summary to make it acceptable
  - » Measure time to post-edit
  - » Also identify accuracy problems in NLP

High Quality Human Eval

- Moramarco et al (2021) describes first version of protocol
- Refined since
  - » Post-edit UI is critical

Correlation with exist metrics

- Preliminary work, not yet published
- Levenshtein (character edit distance) better than ROUGE, BertScore, etc

Metric	Corr with post-edit time	Corr with num error
ROUGE-2	0.38	0.73
METEOR	0.41	0.71
BertScore	0.50	0.74
Levenshtein	0.55	0.76

Levenshtein is best?

- Surprising that Levenstein (character level edit distance is best)
- Bertscore, etc, mostly justified by corr with crowdworker opinion (eg, WMT)
  - » Freitag: Corr between WMT and prof translators can be negative…
  - » Good corr with WMT not guarantee good corr with high-quality outcome-based human evals!

Summary

- Working towards high quality eval of real-world utility
  - » Work in progress
  - » Expensive (need lots of doctor time)
- Explore which metrics have best corr
  - » So far 1960s Leven dist beats all of the modern metrics used in NLP

Contents

- High quality human evaluation
- Old work
- Detecting accuracy errors
- Evaluating real-world utility of summaries
- *Enhancing replicability*
- Final thoughtrs

Reproducibility

- Scientific experiments (including eval of AI systems) should be reproducible!
- If someone else does the same exper, should get similar results
  - » Not identical if people are involved
- Major concern in many areas of science

Reproducibility in NLP

- Some work on reproducing automatic (metric) evals
  - » Ensure all details published, data sets and soft available, preprocessing clear, etc
- What about reproducing human eval?
  - » Poorly understood

ReproGen: Human NLG Eval

- Shared task where people reproduced human evaluations of NLG systems
  - » Belz et al (2021)
- Mixed results
  - » Some reproductions had similar results, some did not
  - » Unclear why (small sample size) (4 replic)

ReproHum

- New EPSRC project on reproducibility of human evaluations of NLP
  - » Will start in early 2022
- Much larger scale than ReproGen
  - » 20 partner labs will reproduce a selected set of NLP evaluations
  - » Identify key factors for replication
  - » Develop theoretical framework
  - » Make recommendations

ReproHum

- New partner labs are welcome!
- Contact Anya Belz (PI) or me if interested

Contents

- High quality human evaluation
- Old work
- Detecting accuracy errors
- Evaluating real-world utility of summaries
- Enhancing replicability
- *Final thoughts*

Final Thoughts

- Too much focus on quick/cheap evals in NLP!
- If we're doing science (as opposed to keeping score in contests), we need high-quality human evals
  - » Ground/validate metrics
  - » Confidence in key findings

Final Thoughts

- I’d love to see more high-quality human evaluations in NLP
- Feel free to contact me if I can help!

References

Belz et al (2021). The ReproGen Shared Task on Reproducibility of Human Evaluations in NLG: Overview and Results. *Proc of INLG-2021*

Freitag et al (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. Arxiv

Garneau and Lamontagne (2021). Shared Task in Evaluating Accuracy: Leveraging Pre-Annotations in the Validation Process. *Proc of INLG-2021*

Hunter et al (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine* **56**:157–172

Kasner et al (2021). Text-in-Context: Token-Level Error Detection for Table-to-Text Generation. *Proc of INLG-2021*

References

Law et al (2005). A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit *Journal of Clinical Monitoring and Computing* **19**:183-94.

Moramarco et al (2021). A preliminary study on evaluating Consultation Notes with Post-Editing. *Proc of EACL-2021 workshop on Human Evaluation of NLP Systems*

Portet et al (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence* **173**:789-816

Reiter et al (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* **144**:41-58

References

Reiter (2018). A Structured Review of the Validity of BLEU. *Computational Linguistics* **44**:393-401

Thomson and Reiter (2020). A Gold Standard Methodology for Evaluating Accuracy in Data-To-Text Systems. *Proc of INLG-2020*

Thomson and Reiter (2021). Generation Challenges: Results of the Accuracy Evaluation Shared Task. *Proc of INLG-2021*