NLG Evaluation

Ehud Reiter

(Abdn Uni and Arria/Data2text)

Structure

- Evaluation Concepts
- Specifics: controlled ratings-based eval

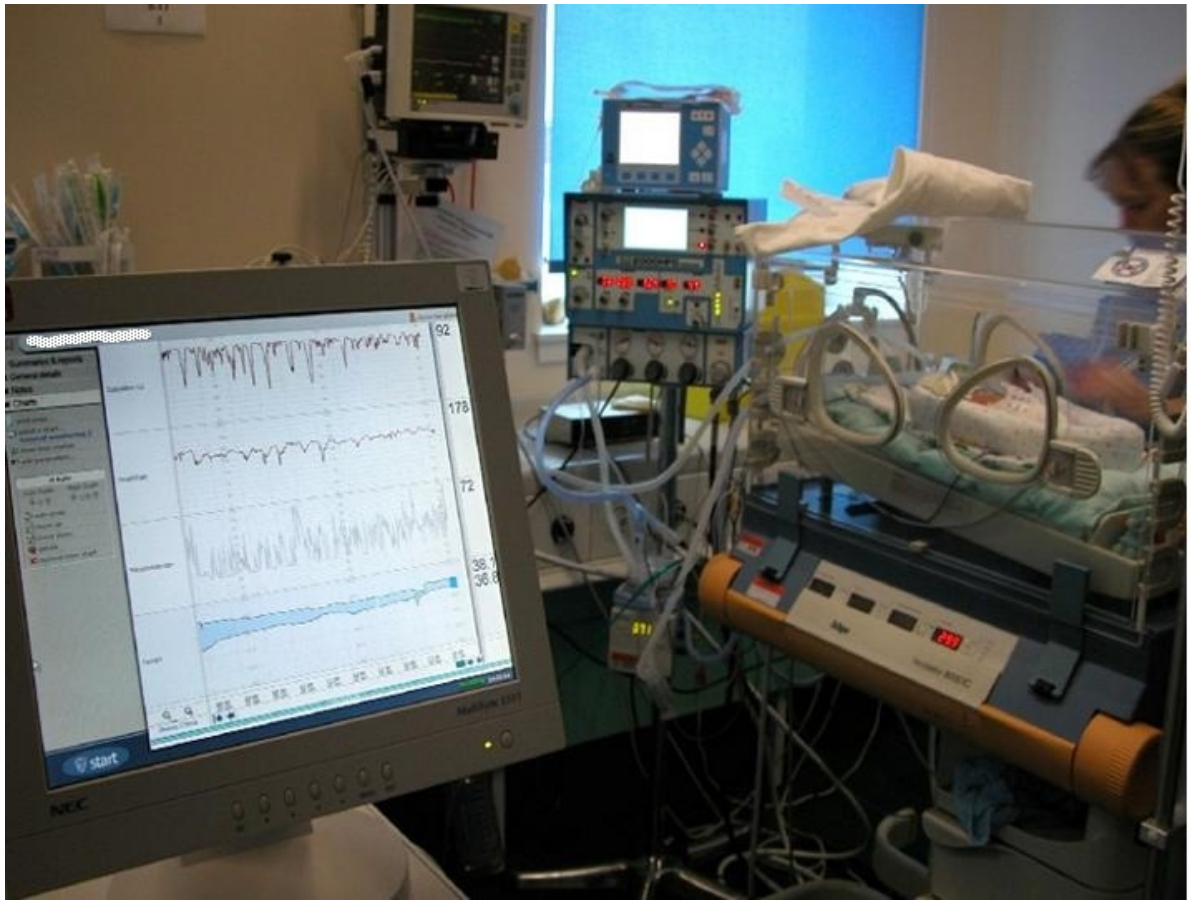
Purpose of Evaluation

- What do we want to know?
- *Audience?*

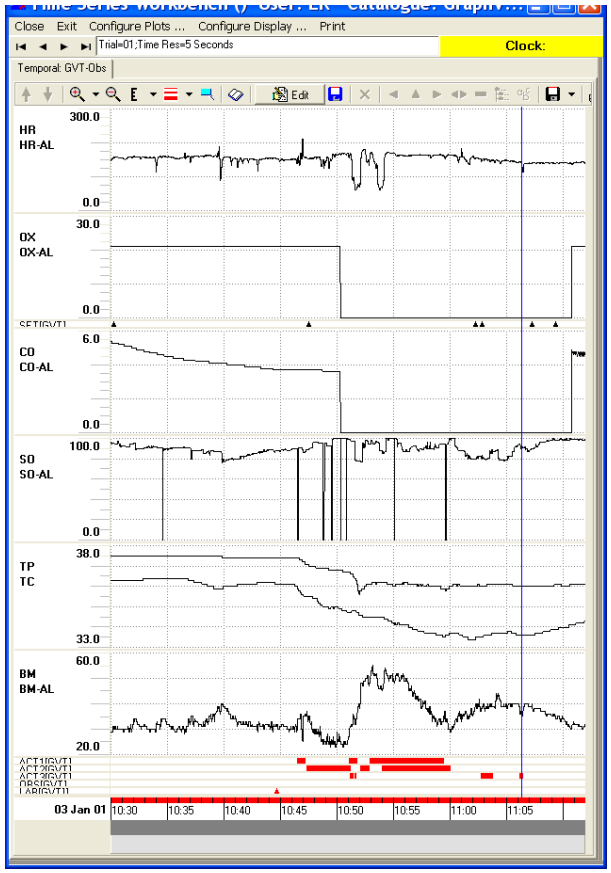
Example: BabyTalk

- Goal: Summarise clinical data about premature babies in neonatal ICU
- Input: sensor data; records of actions/ observations by medical staff
- Output: multi-para texts, summarise
  - » BT45: 45 mins data, for doctors
  - » BT-Nurse: 12 hrs data, for nurses
  - » BT-Family: 24 hrs data, for parents

Babytalk: Neonatal ICU



Babytalk Input: Sensor Data



BT-Nurse text (extract)

Respiratory Support

Current Status

Currently, the baby is on CMV in 27 % O2. Vent RR is 55 breaths per minute. Pressures are 20/4 cms H2O. Tidal volume is 1.5.

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.

…

Babytalk eval: goals

- Babytalk evaluation goals
  - » Medics want to know if Babytalk summaries enhance patient outcome
    - – Deploy Babytalk on ward and measure outcome (RCT)
  - » Psychologists want to know if Babytalk texts are effective decision support tool
    - – Controlled “off ward” study of decision effectiveness
  - » Developers want to know how improve system
    - – Qualitative feedback often most useful
  - » Software house wants to know if profitable
    - – Business model (costs and revenue)

Which Goal?

- Depends!
  - » Publish NLG research papers – usually focus on “psychologist” goals
  - » Publish NLP research paper – usually performance on standard data set
    - – Very dubious in my opinion….
- But other goals also important

Types of NLG Evaluation

- Task Performance
- Human Ratings
- Metric (comparison to gold standard)

- Controlled vs Real-World

Task-Performance Eval

- Measure whether NLG system achieves its communicative goal
  - » Typically helping user perform a task
  - » Other possibilities, eg behaviour change
- Evaluate in real-world or in controlled experiment

Real world: STOP smoking

- STOP system generates personalised smoking-cessation letters
- Recruited 2553 smokers
  - » Sent 1/3 STOP letters
  - » Sent 1/3 fixed (non-tailored) letter
  - » Sent 1/3 simple “thank you” letter
- Waited 6 months, and compared smoking cessation rates between the groups

Results: STOP

- 6-Month cessation rate
  - » STOP letter: 3.5%
  - » Non-tailored letter: 4.4%
  - » Thank-you letter: 2.6%
- Note:
  - » More heavy smokers in STOP group
  - » Heavy smokers less likely to quit

Negative result

- Should be published!
- Don't ignore or "tweak stats" until you get the "right" answer

- E Reiter, R Robertson, and L Osman (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* **144**:41-58.

Controlled exper: BT45

- Babytalk BT-45 system (short reports)
- Choose 24 data sets (scenarios)
  - » From historical data (5 years old)
- Created 3 presentations of each scenario
  - » BT45 text, Human text, Visualisation
- Asked 35 subjects (medics) to look at presentations and decide on intervention
  - » In experiment room, not in ward
  - » Compared intervention to gold standard
- Computed likelihood of correct decision

Results: BT45

- Correct decision made
  - » BT45 text: 34%
  - » Human text: 39%
  - » Visualisation: 33%
- Note:
  - » BT45 texts mostly as good as human, but were pretty bad in scen where target action was “no action” or “sensor error”

Reference

- F Portet, E Reiter, A Gatt, J Hunter, S Sripada, Y Freer, C Sykes (2009). Automatic Generation of Textual Summaries from Neonatal Intensive Care Data. *Artificial Intelligence* **173**:789-816
- M. van der Meulen, R. Logie, Y. Freer, C. Sykes, N. McIntosh, and J. Hunter, "When a graph is poorer than 100 words: A comparison of computerised natural language generation, human generated descriptions and graphical displays in neonatal intensive care," *Applied Cognitive Psychology*, vol. 24, pp. 77-89, 2008.

Task-based evaluations

- Most respected
  - » Especially outwith NLG/NLP community
- Very expensive and time-consuming
- Eval is of specific system, not generic algorithm or idea
  - » Small changes to both STOP and BT45 would probably changes eval result

Human Ratings

- Ask human subjects to assess texts
  - » Readability (linguistic quality)
  - » Accuracy (content quality)
  - » Usefulness
- Can assess control/baseline as well
- Usually use Likert scale
  - » Strongly agree, agree, undecided, disagree, strongly disagree (5 pt scale)

Real world: BT-Nurse

- Deployed BT-Nurse on ward
- Nurses used it on real patients
  - » Both beginning and end of shift
  - » Vetted to remove content that could damage care
    - – No content removed
- Nurses gave scores (3-pt scale) on each text
  - » Understandable, accurate, helpful
  - » Agree, neutral, disagree
- Also free-text comments

Results: BT-Nurse

- Numerical results
  - » 90% of texts understandable
  - » 70% of texts accurate
  - » 60% of texts helpful
  - » [no texts damaged care]
- Many comments
  - » More content
  - » Software bugs
  - » A few “really helped me” comments

Reference

- J Hunter, Y Freer, A Gatt, E Reiter, S Sripada, C Sykes, D Westwater (2011). BT-Nurse: Computer Generation of Natural Language Shift Summaries from Complex Heterogeneous Medical Data. *Journal of the American Medical Informatics Association* **18**:621-624
- J Hunter, Y Freer, A Gatt, E Reiter, S Sripada, C Sykes (2012). Automatic generation of natural language nursing shift summaries in neonatal intensive care: BT-Nurse. *Artificial Intelligence in Medicine* **56**:157–172

Controlled exper: Sumtime

- Marine weather forecasts
- Choose 5 weather data sets (scenarios)
- Created 3 presentations of each scenario
  - » Sumtime text
  - » human texts
  - » Hybrid: Human content, SumTime language
- Asked 73 subjects (readers of marine forecasts) to give preference
  - » Each saw 2 of the 3 possible variants of a scenario
  - » Most readable, most accurate, most appropriate

Results: SumTime

Question	ST	Human	same	p value
SumTime vs. human texts				
More appropriate?	43%	27%	30%	0.021
More accurate?	51%	33%	15%	0.011
Easier to read?	41%	36%	23%	>0.1
Hybrid vs. human texts				
More appropriate?	38%	28%	34%	>0.1
More accurate?	45%	36%	19%	>0.1
Easier to read?	51%	17%	33%	<0.0001

Reference

- E Reiter, S Sripada, J Hunter, J Yu, and I Davy (2005). Choosing Words in Computer-Generated Weather Forecasts. *Artificial Intelligence* **167**:137-169.

Human ratings evaluation

- Probably most common type in NLG
  - » Well accepted in academic literature
- Easier/quicker than task-based
  - » For controlled eval, can be able to use Mechanical Turk
  - » Can answer questions which are hard to fit into a task-based evaluation
    - – Can ask people to generalise

Metric-based evaluation

- Create a gold standard set
  - » Input data for NLG system (scenarios)
  - » Desired output text (usually human-written)
    - – Sometimes multiple “reference” texts specified
- Run NLG system on above data sets
- Compare output to gold standard output
  - » Various metrics, such as BLEU
- Widely used in machine translation

Example: SumTime input data

day/hour	wind-dir	speed	gust
05/06	SSW	18	22
05/09	S	16	20
05/12	S	14	17
05/15	S	14	17
05/18	SSE	12	15
05/21	SSE	10	12
06/00	VAR	6	7

Example: Gold standard

- Reference 1 - SSW'LY 16-20 GRADUALLY BACKING SSE'LY THEN DECREASING VARIABLE 4-8 BY LATE EVENING
- Reference 2 - SSW 16-20 GRADUALLY BACKING SSE BY 1800 THEN FALLING VARIABLE 4-8 BY LATE EVENING
- Reference 3 - SSW 16-20 GRADUALLY BACKING SSE THEN FALLING VARIABLE 04-08 BY LATE EVENING

Above written by three professional forecasters

Metric evaluation example

- SumTime output:
  - » SSW 16-20 GRADUALLY BACKING SSE THEN BECOMING VARIABLE 10 OR LESS BY MIDNIGHT
- Compare to Reference 1
  - » SSW~~'LY~~ 16-20 GRADUALLY BACKING SSE~~'LY~~ THEN ~~DECREASING~~ BECOMING VARIABLE ~~4-8~~ 10 OR LESS BY ~~LATE EVENING~~ MIDNIGHT
- Compute score using metric
  - » edit distance, BLEU, etc

Issues

- Is SSW'LY better than SSW?
  - » 2 out of 3 reference texts use SSW
  - » Good to have multiple reference texts
- Is BY LATE EVENING better than BY MIDNIGHT?
  - » User studies with forecast readers suggest BY MIDNIGHT is less ambiguous
  - » Should ST be eval against human texts?
    - – SumTime texts are better than human texts!

Which Metric is Best?

- Assess by validation study
  - » Do "gold standard" eval of multiple systems
    - Task-performance or human ratings
    - Ideally evaluate 10 or more NLG systems
      - Which must have same inputs and target outputs
  - » Also evaluate systems using metrics
  - » Which metric correlates best with "gold standard" human evaluations?
    - Do any metrics correlate?

Validation: result

- NIST-5 (BLEU variant) is best predictor of human clarity (readability) judgements
- No metric correlates with human accuracy judgements
- E Reiter, A Belz (2009). An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems *Computational Linguistics* **35**:529–558

Metric-based evaluation

- I seriously dislike
  - » we don't have strong evidence that metrics predict human ratings, let alone task perf
  - » Also people can "game" the metrics
- (my opinion) have distorted machine translation, summarisation
  - » Communities forced to use poorly validated metrics for political/funding reasons
  - » Not the way to do good science

General issues

- Validity
  - » Are eval technique correlated with goal?
    - – Do human ratings correlate with task performance?
    - – BT: subjects did best with human text summaries, but preferred the visualisations
  - » Psych: why do US universities use SAT?
- Generalisability
  - » Do results generalise (domains, genres, etc)?
    - – ST: Can we generate good aviation forecasts?
  - » Psych: intelligence tests don't work on minorities

Statistics

- Be rigorous!
  - » Non-parametric tests where appropriate
  - » Multiple hypothesis corrections
  - » Two-tailed p-values
  - » Avoid post-hoc analyses
- Medicine: strict stats needed
  - » Are “significant” results replicated?
  - » Only if stats are very rigorous

Specifics

- How perform a controlled ratings-based evaluation?
- Example: weather forecasts

Experimental Design

- Hypotheses
- Subjects
- Material
- Procedure
- Analysis

Hypotheses: before experiment

- Define hypotheses, stats, etc in detail before the experiment is done
  - » In medicine, expected to publish full experimental design beforehand
    - – https://clinicaltrials.gov/
  - » If multiple hypothesis, reduce p value for significance (discuss later)
- Why?

Post-hoc

- Colleague once told me “I didn’t see a significant effect initially, so I just loaded the data into SPSS and tried all kinds of stuff until I saw something with $p < .05$”
- What is wrong with this?

Why is this bad?

- Assume we test 10 variants of a hypothesis
  - » “Sumtime more accurate than human”
  - » “Hybrid more readable than human”
  - » etc
- Assume we use 10 different stat tests
  - » Eg, normalise data in different ways
- 100 tests
  - » so we’ll see a (variant, stat) combination which is sig at p = .01, even if no genuine effect

Hypotheses: SumTime

- Hyp 1: Sumtime texts more appropriate than human texts
- Hyp 2: Hybrid texts more readable than human texts

- 2 hypotheses, so significant of $p < .025$
- Any other hypothesis post-hoc
  - » Including “ST texts more accurate”
  - » Not significant even though $p = 0.011$

Subjects: Who are they

- What subjects are needed
  - » Language skills? Domain knowledge? Background? Age? Etc
- Sometimes not very restrictive
  - » General hypotheses about language
  - » Mechanical Turk is good option
- Sometimes want specific people
  - » Eg, test reaction of users to a system
    - – Babytalk-Family evaluated by parents who have babies in neonatal ICU

How many subjects?

- Can do a *power calculation* to determine subject numbers
  - » https://en.wikipedia.org/wiki/Statistical_power
  - » Depends on expected effect size
    - – More subjects needed for smaller effects
  - » Typically looking for 50+
    - – Not a problem with Mturk
    - – Can be real hassle if need subjects with specialised skills or backgrounds

Recruitment of subjects

- General subjects (easier)
  - » Mechanical Turk
  - » Local students
- Specialised subjects (harder)
  - » email lists, networks, conferences, …
  - » Personal contacts

Subjects: SumTime

- Type: regular readers and users of marine weather forecasts
- Recruitment: asked domain experts (working on project) to recruit via their networks and contacts
- Number: wanted 50, got 72

Material: scenarios

- Usually start by choosing some scenarios (data sets)
  - » Usually try to be representative and/or cover important cases
  - » Random choice also possible

Material: presentations

- Typically prepare different presentations of each scenario
  - » Output of NLG system(s) being evaluated
  - » Control/baseline
    - – Human-authored text?
    - – Output of current best-performing NLG system?
    - – Fixed (non-generated) text?
  - » Depends on hypotheses

Material: structure

- For each scenario, subjects can see
  - » One presentation
  - » Some presentations
  - » All presentations
- Subjects should not know if a presentation is NLG or control!

Material: Sumtime

- Number of scenarios: 5
  - » Corpus texts written by 5 forecasters
  - » First text written by forecaster after magic date
  - » Wanted human/control texts from each of 5
- Number of presentations: 3
  - » SumTime (main)
  - » Human (control)
  - » Hybrid (of content-det vs microplan/real)
- Structure:
  - » Present pairs (2 out of 3) to each subject

Procedure: What subject do

- What questions asked
  - » Readable, accurate, useful
  - » Response: N-pt Likert scale, slider
    - – https://en.wikipedia.org/wiki/Likert_scale
- Order
  - » Latin Square (Balanced)
  - » Random
- Payment?

Latin Square

	Scenario 1	Scenario 2	Scenario 3
Subject 1	SumTime	Human	Hybrid
Subject 2	Hybrid	SumTime	Human
Subject 3	Human	Hybrid	SumTime

Procedure: Questions

- Practice questions at beginning?
- Fillers between questions we care about?
- Especially important if we want timings

Procedure: Ethics

- Can doing experiment harm people?
  - » BT-Nurse and patient care
  - » If so, must present acceptable solution
- Subjects can drop out at any time
  - » Can NOT “pressure” them to stay if the want to quit experiment

Procedure: Exclusion

- When do we drop a subject from the experiment?
  - » Incomplete responses?
  - » Inconsistent responses?
  - » Bizarre responses?

Procedure: SumTime

- Questions
  - » Presented 2 variants
  - » Which variant is: easiest to read; most accurate; most appropriate
- Order not randomised
- No payment
- No practice or filler, no ethical issues
- Excluded if less than 50% completed

Statistics: Test

- Principle: Likert scales are not numbers
  - » Should not be averaged
  - » Non-parametric test (Wilcoxon Signed Rank)
- Practice
  - » Often present average Likert score
  - » Use parametric test, such as t-test
  - » More or less works….
    - – But not if rigorous stats needed!

Statistics: Normalisation

- Some users are more generous than others
- Some scenarios are harder than others
- Potential bias
  - » User X always “agree”, Y always “disagree”
  - » X rates 10 SumTime texts and 1 corpus text
  - » Y rates 1 SumTime text and 10 corpus texts
- Use balanced design (Latin square)
- Use linear model
  - » Predicts score on user, scenario, presentation
  - » Just look at presentation element

Statistics: Multiple Hypoth

- Bonferroni multiple hypothesis correction
- Divide significance p value by number of hypotheses being tested
  - » 1 hypothesis: look for $p < .05$
  - » 2 hypotheses: look for $p < 0.025$
  - » 10 hypotheses: look for $p > 0.005$

Statistics: SumTime

- Test: Chi-square
  - » Because users asked to state a preference between variants, did not give Likert score
- Normalisation: not necessary
  - » Less important with preferences
    - – If user is asked whether A or B is better, doesnt matter how generous he is (“great” vs “poor”)
- Multiple hypotheses: $p < 0.025$
  - » Because 2 hypotheses

Questions