

NLG EVALUATION

Ehud Reiter

<http://ehudreiter.com>

Sept 2017



ARRIA
NATURAL LANGUAGE GENERATION

Contents

- Background: Evaluation and NLG
- Experimental design
- Types of NLG Evaluation: Task, Ratings, Metric
- Commercial evaluation
- Case Studies
- Q&A

Audience

- I assume attendees are familiar with NLG
- I assume that most attendees have some experience evaluating NLG systems
 - Or at least have read about NLG evaluations
- I also assume that attendees are not experts at NLG evaluation

Background: Evaluation and NLG

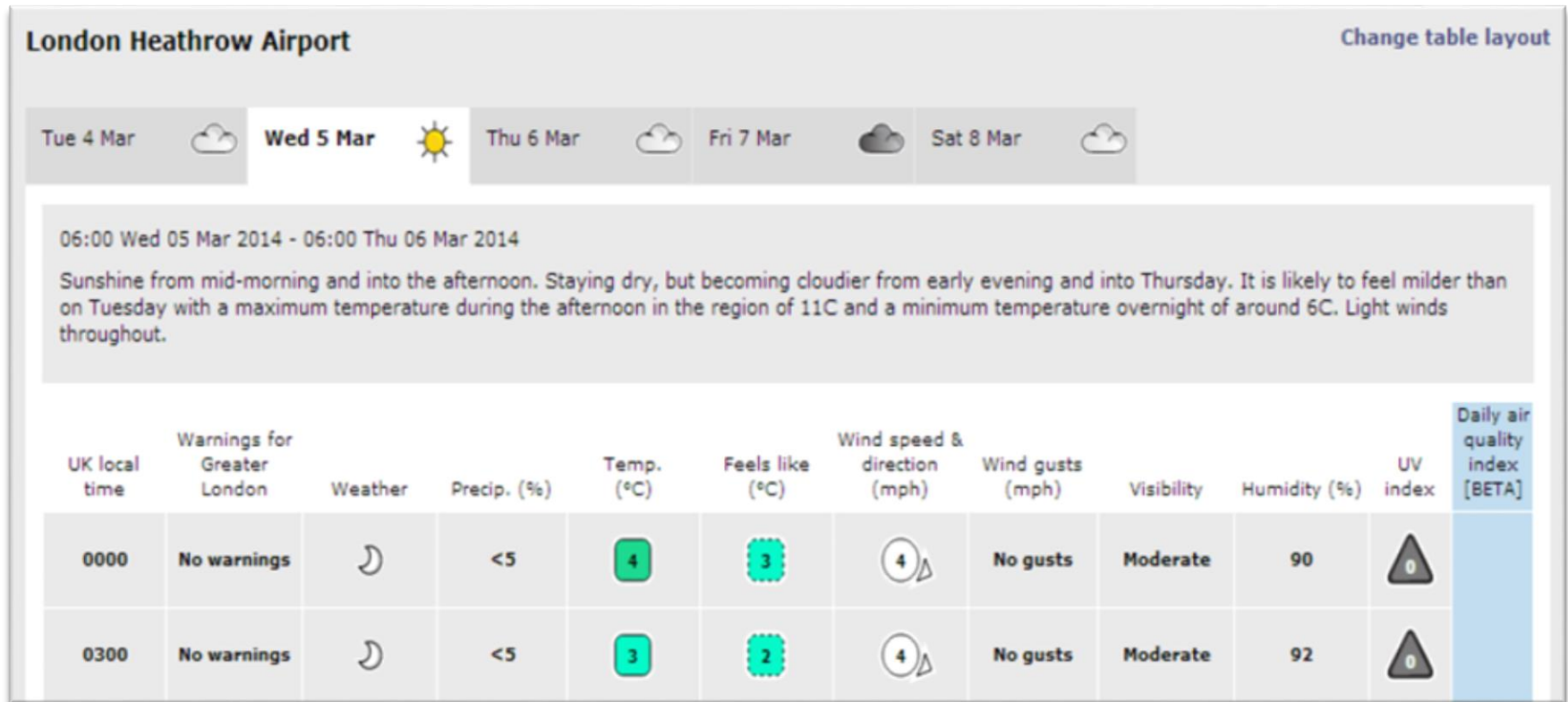
What is evaluation?

- Experimentally testing hypotheses about performance
 - Is system/algorithm/model/etc X better than baseline or state-of-the-art?
 - Is system/algorithm/model/etc. X useful in real-world applications?
- Of course there are many other kinds of hypothesis which we can test
 - Eg, cognitive modelling

Natural Language Generation

- Software which generates texts in English (French, etc.) from semantic representations and/or non-linguistic data
 - Textual weather forecasts from numerical weather prediction models
 - Summary for patient from electronic patient record
 - Financial reports from finance spreadsheets
 - Etc.

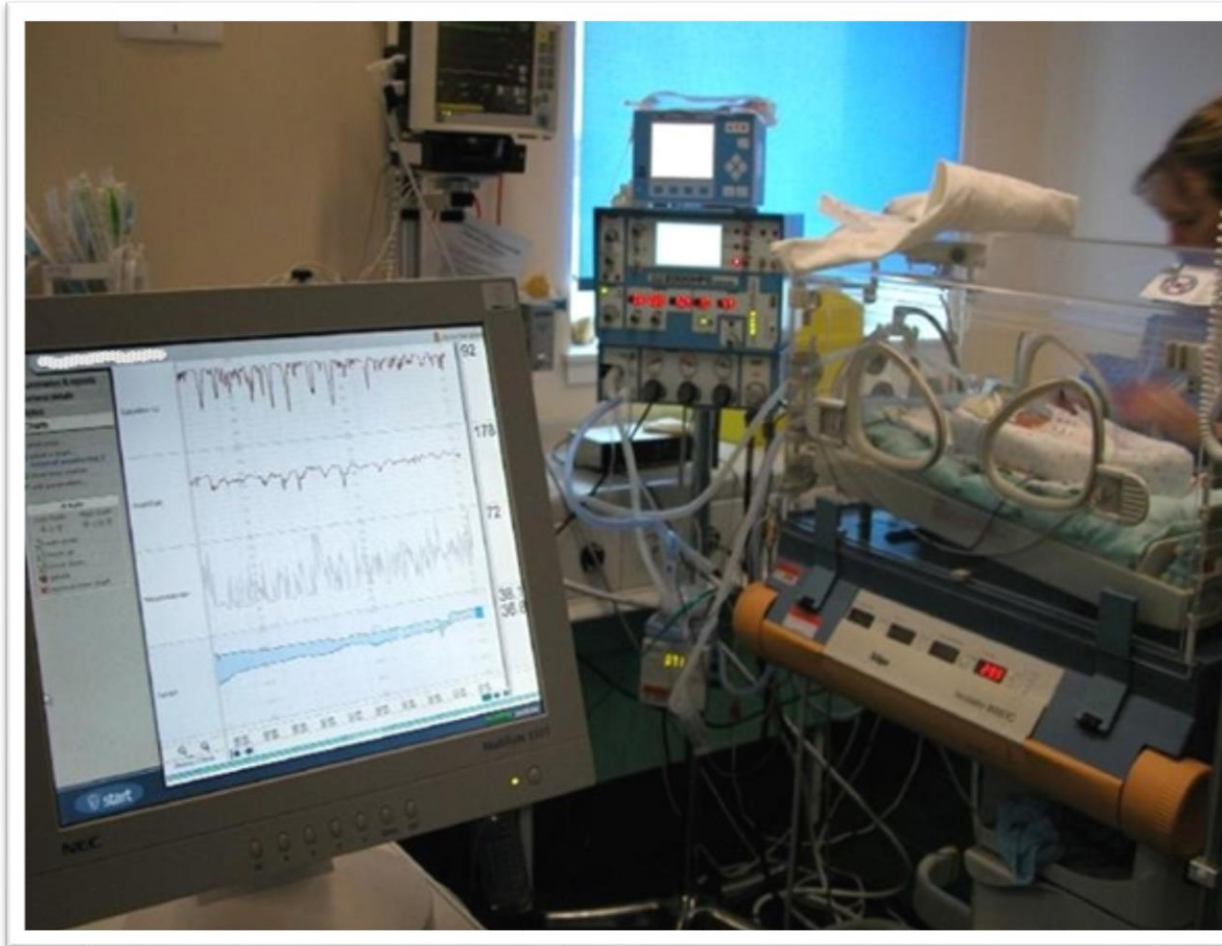
Simple example: Point weather forecast



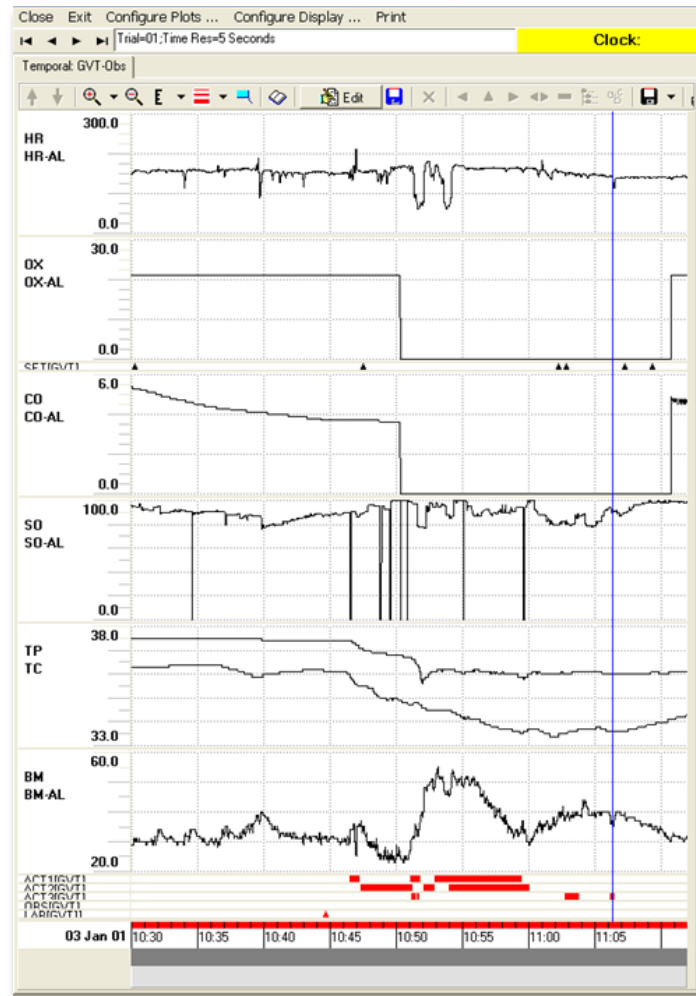
Complex: BabyTalk

- Summarised clinical data about premature babies in neonatal ICU
- Input: sensor data; records of actions and observations by medical staff
- Output: multi-paragraph texts, summarised data for different audiences

Babytalk: Neonatal ICU



Babytalk Input: Sensor Data



Input: Action Records

Full Descriptor	Time
SETTING;VENTILATOR;FiO2 (36%)	10.30
MEDICATION;Morphine	10.44
ACTION;CARE;TURN/CHANGE POSITION;SUPINE	10.46 - 10.47
ACTION;RESPIRATION;HAND-BAG BABY	10.47 - 10.51
SETTING;VENTILATOR;FiO2 (60%)	10.47
ACTION;RESPIRATION;INTUBATE	10.51 - 10.52

BT45 texts (extract)

- Short summary supporting real-time decision making by clinicians

By 11:00 the baby had been hand-bagged a number of times causing 2 successive bradycardias. She was successfully re-intubated after 2 attempts. The baby was sucked out twice. At 11:02 FIO2 was raised to 79%.

BT-Family text (extract)

- Page-long text for parents

Yesterday, John was on a ventilator. The mode of ventilation is Bilevel Positive Airway Pressure (BiPAP) Ventilation. This machine helps to provide the support that enables him to breathe more comfortably. Since last week, his inspired Oxygen (FiO_2) was lowered from 56 % to 21 % (which is the same as normal air). This is a positive development for your child.

During the day, Nurse Johnson looked after your baby. Nurse Stevens cared for your baby during the night.

BT-Nurse text (extract)

- 5 page shift handover report for nurses

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.

...

Babytalk evaluations

- Different groups interested in different hypotheses and evaluations!
 - 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
 - Software house wants to know if profitable
 - Cost, revenue, risks
 - CS/NLP people want to know how improve system
 - Qualitative feedback often most useful

Experimental Design

Experimental Design

When designing an experiment (for evaluation or otherwise), we need to decide on

- Hypotheses
- Subjects
- Material
- Procedure
- Analysis

Get the basics right!

Hypotheses

- What hypotheses are we testing?
- What do we expect/hope to find
 - Eg, Users will prefer weather forecasts produced by SuperMet system over forecasts produced by FGen system

Hypotheses in Evaluation Experiments

- Evaluation hypotheses are usually about
 - *Utility*: System will achieve some effect, such as behaviour change or improved decision making
 - *User satisfaction*: Users will like and be satisfied by the system
 - *Similar to humans*: System produces texts which are similar to what human writers produce
- Other possibilities, eg compute speed (less common)

Hypotheses for weather forecasts

- *Utility*: improved decision making
 - Weather game: where send ice cream van?
- *User satisfaction*:
 - Ask users to rate satisfaction on Likert scale
 - Ask users which forecasts they prefer
- *Similar to humans*
 - Compare to human forecast using BLEU (etc)

Hypotheses Before Experiment

- Decide on hypotheses ***before*** you do the experiment!!!
 - Write them down somewhere
- Don't change/tweak hypotheses to better fit the data
 - “Results weren't good, so I loaded data into SPSS and played around until I found something significant”
 - Invalidates statistical significance tests
- If data is better fit to an alternative hypothesis
 - report as a post-hoc insight, or do another experiment to test this hypothesis.
 - Don't say alternative hypothesis supported by data

Example

- Hypothesis: Users prefer weather forecasts produced by SuperMet system over forecasts from FGen system
- Result
 - No difference in preference overall
 - But data suggests that women prefer SuperMet, and men prefer FGen
- DON'T say your experiment clearly shows a gender effect
- DO say posthoc analysis suggests may be a gender effect
- DO run another experiment to explicitly test gender effect
 - If this exper successful, can say shown gender effect

Experimental Design

When designing an experiment, we need to decide on

- Hypotheses
- **Subjects**
- Material
- Procedure
- Analysis

Subjects

- If we're doing experiments on people, who are they?
- How many subjects are we looking for?
- How do we recruit them?

Who are Subjects?

- Subjects should be potential users
 - If SuperGen generates marine weather forecasts aimed at sailors, don't ask students to evaluate it!
 - If SuperGen generates public forecasts, anyone can evaluate it, including students
- Ideally subjects should be representative
 - age, gender, expertise, etc
 - Can be hard to achieve in practice
 - You should report subject characteristics
 - Eg, students between ages of 18-25, 45% female

How Many Subjects?

- Ideally numbers based on a statistical power calculation
 - Algorithm/app which tells you numbers based on your experiment and expected effective size
 - Eg, SUPERMET Likert ratings are expected to be 0.5 higher than FGEN ratings
 - Often dont know expected effect size, stan dev
- As a very rough rule of thumb, 50 subjects is often a good number in NLG evaluations
 - Of course depends on the experiment!!

Power calculation (PS)

Power and Sample Size Program: Main Window

File Edit Log Help

Survival t-test Regression 1 Regression 2 Dichotomous Mantel-Haenszel Log

[Studies that are analyzed by t-tests](#)

Output

[What do you want to know?](#) Sample size

[Sample Size](#) 64

Design

[Paired or independent?](#) Independent

Input

α	.05	δ	.5
		σ	1
power	.8	m	1

Calculate

Graphs

Subject Recruitment

- Easiest way to recruit is Mechanical Turk or similar
 - Amazon service where you can hire random people to do small tasks very cheaply.
 - “Task” is participating in your experiment
 - Many alternatives to Amazon
 - Works *if* Turkers are potential users, real-world context isn’t needed, and you don’t need to observe or debrief subjects

Subject Recruitment

- Can recruit friends, colleagues, social network
 - Free (unlike Mturk)
 - Are they potential users, etc?
 - Can you get enough subjects?
 - Will they be biased because they know the outcome you hope to achieve?
 - Do they know SuperMet is your creation?

Subject Recruitment

- Explicit recruitment of subjects
 - Target exactly the kind of subject you want
 - Via contacts, bulletin boards, advertisements, professional conferences, ...
 - Often takes a lot of time and effort...

Experimental Design

When designing an experiment, we need to decide on






- Hypotheses
- Subjects
- **Material**
- Procedure
- Analysis

Material

- Usually our experiments are based on scenarios, which are defined by a set of input data
 - For each scenario, we produce an output text from the system we are evaluating and from the controls
- What systems do we look at?
- How do we choose scenarios?
 - Random choice? Try to cover varied data sets?
- How many scenarios do we need?

Simple scenario







London Heathrow Airport Change table layout

Tue 4 Mar  **Wed 5 Mar**  Thu 6 Mar  Fri 7 Mar  Sat 8 Mar 

06:00 Wed 05 Mar 2014 - 06:00 Thu 06 Mar 2014

Output text

Sunshine from mid-morning and into the afternoon. Staying dry, but becoming cloudier from early evening and into Thursday. It is likely to feel milder than on Tuesday with a maximum temperature during the afternoon in the region of 11C and a minimum temperature overnight of around 6C. Light winds throughout.

UK local time	Warnings for Greater London	Weather	Precip. (%)	Temp. (°C)	Feels like (°C)	Wind speed & direction (mph)	Wind gusts (mph)	Visibility	Humidity (%)	UV index	Daily air quality index [BETA]
0000	No warnings		<5	4	3	4 	No gusts	Moderate	90	0 	
0300	No warnings		<5	3	2	4 	No gusts	Moderate	92	0 	

Input data

Which Systems?

- Usually want to compare our NLG system to other sys
- Which comparison systems do we use?
 - Baseline: ideally state-of-the-art existing system
 - Eg, compare SuperMet to best existing weather forecast generator (FGen)
 - Topline: good-quality human texts?
- Create scenario texts for each system
 - Eg, create SuperGen, FGen, human text for each scenario data set

How do we choose scenarios?

- Easiest approach is to randomly select data sets
- But this means that we'll have lots of examples of common cases, and few examples of unusual cases
 - Handling edge cases is important for quality
 - Eg, SuperGen when there are gale-force winds
 - Unlikely to be in randomly selected data set
- Alternative is to look for diverse data sets, including important edge cases
 - Ie, explicitly include gale-force wind scenario(s)

Which Scenarios?

- Example strategy for evaluating SuperGen: Choose
 - 4 forecasts for boring days
 - 4 forecasts for days with lots of rain
 - 4 forecasts for days with strong winds
 - 8 additional forecasts (random)
- Within each category, choose at random
- Ensures we have coverage of rainy/windy days

How Many Scenarios?

- As many as possible, provided that all of the scenarios are seen an appropriate number of times.
 - Eg, if 50 subjects each look at 4 scenarios, and we want each scenario seen by 10 subjects, then we should have 20 scenarios
- As a very rough rule of thumb, many experiments do use on the order of 20 scenarios
 - Of course depends on the experiment!!

Experimental Design

When designing an experiment, we need to decide on

- Hypotheses
- Subjects
- Material
- **Procedure**
- Analysis

Procedure

- What do we ask subjects to do with the material, and what do we measure?
- High-level choices
 - Human evaluation vs automatic metric
 - Task/outcome based vs ratings based
 - Artificial lab context vs real-world context
 - Discussed later

Which texts do Subjects See?

- Lets say we have 20 scenarios, and three texts for each
 - Baseline, system, human-written topline
 - 60 texts in all
- Assume we have 60 subjects, which look at 3 texts each
- What does each subject see?
 - Randomly choose from 60 texts?
 - Subj 31 sees 7-baseline, 15-baseline, 16-system
 - Show each subject all 3 variants of one text?
 - Subj 31 sees 7-baseline, 7-system, 7-topline
 - *Latin square* – balance design (my favourite)

Latin Square

	Scenario 1	Scenario 2	Scenario 3
Subject 1	SuperMet	Human	FGen
Subject 2	FGen	SuperMet	Human
Subject 3	Human	FGen	SuperMet

Procedure: Other

- Practice scenarios: Can give subjects some practice scenarios which we do not record
 - Especially useful if we are timing people, since people are usually slower the first time
- Exclusion: May drop subjects who don't seem to be taking the experiment seriously
 - As determined by an objective criteria, such as correct response to trivial question

Procedure: Ethics

- Subjects must give informed consent
- Subjects can drop out at any time
 - Can NOT pressure them to stay if they want to quit
- Data is anonymised
- Can experiment harm someone?
 - Can subject be harmed?
 - Can text hurt someone else (medical decision support)
 - If possible, must present acceptable solution
- Most universities have ethics panels which assess and approve procedures from an ethical perspective

Experimental Design

When designing an experiment, we need to decide on

- Hypotheses
- Subjects
- Material
- Procedure
- **Analysis**

Analysis

We have the data, how do we analyse and report it?

- What statistics do we use
- How do analyse and report free-text comments?
- Error analysis

Statistical Hypothesis Testing

- Quick review, assume people know this
- We test the “null hypothesis” that there is no difference between the systems we are evaluating
- Eg, null hypothesis is that there is no difference in user preference between SuperGen and FGen.
 - Some people prefer SuperGen, but just as many prefer Fgen
- How likely is observed data if null hypothesis is true?

Statistical Hypothesis Testing

- Suppose we ask 10 subjects which they prefer?
- What if 6 of our subjects prefer Supergen
 - Cannot reject null hyp, maybe we just happened to choose 6 people who liked Supergen
- What if 9 of our subjects prefer Supergen
 - This is pretty unlikely if null hyp is true, so can reject
- P values give likelihood of observation if null hyp true
 - Usually say null hyp rejected if $p < .05$

Statistical Hypothesis Testing

- Suppose we have a bag containing lots of white and black marbles.
- Null hypothesis: equal number of white and black marbles
- What if we pick 10 marbles, and 6 of these are black?
 - Cannot reject null hypothesis
- What if we pick 10 marbles, and 9 of these are black?
 - This is pretty unlikely if null hyp is true, so can reject

Statistical Test

- Many different statistical tests!
- However, there are standard ones for common types of hypotheses
 - Described in subsequent slides
- Use these unless you have good reason not to!
 - Which is explained in your write-up

Analysis: comparison of mean

- Is the average (mean) of something (user rating, time to make decision) different between the main system and the control/comparison systems
 - Do users give higher Likert ratings to SuperMet forecasts than to FGen forecasts?
- Use *t-test* when comparing means of two systems
 - Eg, SuperMet text ratings vs FGen text ratings

Comparison of mean

- Use *ANOVA* when comparing means of more than two systems
 - Eg SuperMet vs FGen vs Human ratings
- Supplement with Tukey HSD post-test to identify significant differences between individual pairs
 - ANOVA will say there is a sig diff somewhere.
 - Tukey HSD will say there is a sig diff between SuperMet and FGen, but not between SuperGen and Human.

Comparison of mean

- Use *General Linear Model* to look at multiple factors that influence the mean
 - Eg, build model which predicts rating based on system (eg, SuperGen), subject (Fred, who is very generous in his ratings), and scenario (eg, data set 1, which is difficult to describe in words)
 - Useful to separate impact of system from impact of other factors

Comparison of Means

- *Mann-Whitney* and *Wilcoxon Signed Rank* are non-parametric tests
 - More robust and conservative than t-test
- Only occasionally used in NLP

Comparison of categorical outcome

- Use *chi-square* test to see if users of system X more likely to be in a certain category than users of system Y
- Eg, if we compare two systems for encouraging smoking cessation, are users of the first system more likely to quite than users of the second system?
 - If 5 out of 100 users of my system stopped smoking, but only 4 out of 100 users of other system stopped smoking, is this significant?
 - Categories are {StillSmoking, QuitSmoking}

Correlations

- If we have two measurements, does an increase in the first measurement mean that the second is also likely to increase (or decrease)?
- Eg, if we ask subjects to rate texts on both readability and usefulness, does a higher readability score usually imply a higher usefulness score?
- Two common *correlations*
 - *Pearson* (most common)
 - *Spearman* (non parametric)

Doing Stats Right

- Decide on and write down your statistical analysis before you do the experiment (same as hypotheses)
 - Posthoc “tweaked” analyses must be reported as such!
- If you are testing more than one hypothesis, reduce p value accordingly
 - Eg if testing 100 hypotheses, look for $p < .0005$ instead of $p < .05$
 - $p < .05$ means 5% chance of incorrectly rejected null hyp
 - So if testing 100 hyp, will make 5 mistakes!
- Always report 2-tailed p values

Free-text comments

- Good practice to ask subjects to give free-text comments after an experiment
- Useful for understanding *why* things (did not) work!
 - Ratings just tell us that users liked SuperMet forecasts
 - Comments tell us why they liked SuperMet
- Analysis should include summary of free-text comments
 - No standard methodology for how to do this
 - Repeat particularly insightful comments?
 - Categorise comments (eg, content vs wording), report numbers in each category?

Error Analysis

- Good practice to identify scenarios where a system did poorly, and analyse why this happened
 - Eg, users gave SuperMet poor ratings in a scenario with very high temperatures
 - Free-text comments complained SuperMet forecast did not highlight potentially dangerous temperature
 - SuperMet needs to do a better job of describing high-temperature scenarios and dangers
- No standard methodology

Coda: Concerns about experimental design

- Concerns in wider scientific community about experiments
- Statistically significant results should be repeatable
 - Results reflect underlying truth, not experimental noise
 - So other researchers will see the same thing
- But many “significant” findings are **not** replicable!
 - Medicine/biology: only 6 out of 53 important cancer studies could be replicated (11%)
 - Psychology: 36% success rate in replications
- Can we trust experimental findings? Can we trust science?

What Studies are Replicable?

- Good experimental design
 - Eg, Randomised controlled clinical trial
- Multiple-hypothesis corrections
 - If you test 100 false hypotheses, 5 will seem significant at $p < .05$
- No posthoc tweaking of hypotheses, stats, etc
 - Publish on a website *before* the experiment
- Bad sign: field is “hot”
 - hotness encourage sloppy science

Scientific Culture

- Ioannidis: Replication problems more likely in fields where
 - Few negative results reported
 - Significance-chasing behaviour
 - My first hypothesis was not significant, so I'll just tweak hypothesis and stats until I get a significant result
 - Underpowered studies
- Unfortunately, NLP meets above criteria...

Hypothesis Testing

- Remember that you are testing hypotheses when you evaluate a system
- Remember that poor hypothesis testing (because of poor experimental design) is meaningless.

Types of Evaluation

Task, Human, Metric

Evaluation Procedures

- Task/outcome based
 - Does the NLG system help users perform a task, or change their behaviour?
- Ratings/opinion based
 - Do users like texts produced by the NLG system?
- Metric based
 - Are the NLG texts similar to human-written texts?
- Evaluate in real world or in laboratory experiment

NLG Evaluations

	Task	Ratings	Metric
Real-world	Stop	BT Nurse	NA
Laboratory	BT 45	Sumtime	<i>weather</i>

Task/Outcome Evaluation

- Extrinsic evaluation
- Measure whether NLG system achieves its goal
 - Better decision making
 - Better clinical outcome
 - Behaviour change
 - Etc.
- Evaluate in real world or in laboratory experiment

Real world: STOP smoking

- STOP system generates personalised smoking-cessation letters

Smoking Information for Heather Stewart

You have good reasons to stop...

People stop smoking when they really want to stop. It is encouraging that you have many good reasons for stopping. The scales show the good and bad things about smoking for you. They are tipped in your favour.

THINGS YOU LIKE

It's relaxing
It stops stress
you enjoy it
it relieves boredom
it stops weight gain
it stops you craving



THINGS YOU DISLIKE

It makes you less fit
It's a bad example for kids
you're addicted
It's unpleasant for others
other people disapprove
It's a smelly habit
It's bad for you
It's expensive
It's bad for others' health

You could do it...

Most people who really want to stop eventually succeed. In fact, 10 million people in Britain have stopped smoking - and stayed stopped - in the last 15 years. Many of them found it much easier than they expected.

Although you don't feel confident that you would be able to stop if you were to try, you have several things in your favour.

- You have stopped before for more than a month.
- You have good reasons for stopping smoking.
- You expect support from your family, your friends, and your workmates.

We know that all of these make it more likely that you will be able to stop. Most people who stop smoking for good have more than one attempt.

Overcoming your barriers to stopping...

You said in your questionnaire that you might find it difficult to stop because smoking helps you cope with stress. Many people think that cigarettes help them cope with stress. However, taking a cigarette only makes you feel better for a short while. Most ex-smokers feel calmer and more in control than they did when they were smoking. There are some ideas about coping with stress on the back page of this leaflet.

You also said that you might find it difficult to stop because you would put on weight. A few people do put on some weight. If you did stop smoking, your appetite would improve and you would taste your food much better. Because of this it would be wise to plan in advance so that you're not reaching for the biscuit tin all the time. Remember that putting on weight is an overeating problem, not a no-smoking one. You can tackle it later with diet and exercise.

And finally...

We hope this letter will help you feel more confident about giving up cigarettes. If you have a go, you have a real chance of succeeding.

With best wishes,

The Health Centre.



STOP experimental design

- Hypothesis: STOP users more likely to quite than controls
- Subjects: recruited 2553 smokers
 - Approached all smokers registered with a set of GPs
 - 1/3 agreed to participate
- Material
 - Three “systems”: STOP, fixed (non-tailored) letter, simple “thank you” letter
 - STOP letter personalised for each subject, other letters were fixed

STOP experimental design

- Procedure
 - Subjects fill out smoking questionnaire
 - Subjects receive STOP, non-tailored, or thank-you letter
 - After 6 months, ask subjects if they have stopped smoking
 - Verify with saliva sample
- Statistics: Chi-square to compare cessation rates in the different groups

STOP: result

- 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

Comment: Negative result

- We published as a negative result
- Negative results can and should be published!
 - Ioannidis: lack of negative results is a very bad sign
 - Negative results can be published: STOP result published in ACL, BMJ, AI Journal
- Publish the result of your experiment, even if it isn't the result you were looking for!

References

- E Reiter, R Robertson, and L Osman (2003). Lessons from a Failure: Generating Tailored Smoking Cessation Letters. *Artificial Intelligence* 144:41-58
- AS Lennox, LM Osman, E Reiter, R Robertson, J Friend, I McCann, D Skatun, and P Donnan (2001). The Cost-Effectiveness of Computer-Tailored and Non-Tailored Smoking Cessation Letters in General Practice: A Randomised Controlled Trial. *British Medical Journal* 322:1396-1400

Laboratory task evaluation

- Real-world task evaluation is “gold standard” evaluation in many fields, including medicine
- But expensive, and can raise ethical issues
- Also can be hard to compare alternatives in a controlled way
- Alternative is *laboratory task evaluation*

Laboratory evaluation: BT45

- Babytalk BT-45 provided decision support to clinicians making decisions about care of premature babies in neonatal ICU

By 11:00 the baby had been hand-bagged a number of times causing 2 successive bradycardias. She was successfully re-intubated after 2 attempts. The baby was sucked out twice. At 11:02 FIO2 was raised to 79%.

BT45 experimental design

- Hypothesis: Clinicians make better decisions from BT45 texts than from standard data visualisations
- Subjects: 35 NICU clinicians, junior nurses to senior doctors
 - Balanced for expertise, not age or gender
- Material
 - 24 data sets (scenarios) from historical data
 - 3 examples of 8 conditions
 - Created 3 presentations of each data set (scenario)
 - BT45 text, Human text, Visualisation

BT45 experimental design

- Procedure: Asked clinicians to look at a presentation and choose intervention (in 3 min)
 - In experiment room, not in ward!
 - Compared intervention to gold standard
 - Decided upon by expert clinicians, no time limit
 - Latin square design
- Analysis: Compare mean decision quality between BT45, human, and visualisation conditions (ANOVA)
 - Different measures of decision quality

Results: BT45

- Correct (gold-standard) decision made (simplified scoring)
 - BT45 text: 34%
 - Human text: 39%
 - Visualisation: 33%
- No significant difference between BT45 and Visualisation

Comment: Edge cases matter

- Error analysis: BT45 texts mostly as good as human, but were did poorly when desired intervention was “no action” or “reattach sensors”
- System needs to perform well in all cases
 - “reattach sensor” as well as “increase oxygen level”
 - In NICU, “no action” is probably most common decision, needs to be handled well!

References

- 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, J. Hunter (2008). 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*, 24:77-89

Task/outcome (extrinsic) evaluations

- Most respected
 - Especially in broader scientific community
- Expensive and time-consuming
- Evaluation is of specific system, not generic algorithm or idea
 - Small changes to BT45 (STOP?) would have significantly changed evaluation result

Human Ratings

- Ask human subjects to assess texts
 - Readability (linguistic quality)
 - Accuracy
 - Usefulness
- Often use Likert scale
 - Strongly agree, agree, undecided, disagree, strongly disagree
- Alternative is for subjects to rank texts on above criteria
 - Eg, rank SuperGen, Fgen, Human forecast on usefulness

Why use ratings?

- Task success criteria are hard to measure
 - Humour, entertainment
 - Public weather forecast
 - “ice cream test” is just one use case
 - etc
- Want subjects to focus on generalise beyond a specific systems
- Task evaluation is too expensive, time-consuming, etc

Real world: BT-Nurse

- BT-Nurse generated shift handover reports for NICU nurses

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: Experimental Design

- Hypothesis: NICU nurses will find BT-Nurse texts to be understandable, accurate, and helpful
 - Not measuring medical outcome
- Subjects: NICU 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

BT-Nurse: 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 that rated texts understandable, etc
 - Use chi-square for significance
 - Quantitative summary of free-text comments

BT-Nurse: 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 content
 - A few “really helped me” comments
 - Some comments highlighted software bugs

Comments: Fix the bugs!

- Fix the bugs before you evaluate!
- “Boring” engineering, but you’ll get poor results if you don’t
 - Note BT-Nurse bugs impacted ratings, but did not damage patient care
 - Babytalk was a university research project
 - Arria puts a lot of effort into testing and quality assurance!

References

- 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* 18:621-62
- 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

Laboratory ratings evaluation

- Laboratory study can be cheaper and quicker
- Easier to compare alternatives in a controlled way

Laboratory experiment: Sumtime

Marine weather
forecasts

Section 2. FORECAST 6 - 24 GMT, Wed 12-Jun 2002

Field	Text
WIND(KTS) 10M	W 8-13 backing SW by mid afternoon and S 10-15 by midnight.
WIND(KTS) 50M	W 10-15 backing SW by mid afternoon and S 13-18 by midnight.
WAVES(M) SIG HT	0.5-1.0 mainly SW swell.
WAVES(M) MAX HT	1.0-1.5 mainly SW swell falling 1.0 or less mainly SSW swell by afternoon, then rising 1.0-1.5 by midnight.
WAVE PERIOD (SEC)	Wind wave 2-4 mainly 6 second SW swell.
WINDWAVE PERIOD (SEC)	2-4.
SWELL PERIOD (SEC)	5-7.
WEATHER	Mainly cloudy with light rain showers becoming overcast around midnight.
VISIBILITY (NM)	Greater than 10.
AIR TEMP(C)	8-10 rising 9-11 around midnight.
CLOUD (OKTAS/FT)	4-6 ST/SC 400-600 lifting 6-8 ST/SC 700-900 around midnight.

Sumtime: Experimental Design

- Hypothesis: Users prefer SumTime texts over human texts
- Subjects: 73 people who read marine forecasts
 - Recruited via contacts
- Material
 - Chose 5 weather data sets (scenarios) from historical data
 - Created 3 presentations of each scenario
 - Sumtime text
 - Human texts (actual forecaster text)
 - Hybrid: Human content, SumTime language
 - Want to assess value of microplanning without doc planning

Sumtime: Experimental Design

- Procedure
 - Subjects shown 2 of the 3 possible variants of a scenario
 - Asking which variant was more readable, most accurate, most appropriate
 - Online or on Word document (hard copy or electronic)
 - Sailors could fill out whilst at sea
- Analysis
 - Chi-square to see if preferences significant

Results: Sumtime

SumTime vs. human texts

Question	SumTime	Human	same	p value
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

Question	Hybrid	Human	same	p value
More appropriate?	38%	28%	34%	0.1
More accurate?	45%	36%	19%	0.1
Easier to read?	51%	17%	33%	>0.0001

Comment: Better Than Human!

- NLG systems can produce texts which are better than human texts!
 - I.e. better than texts written by humans of average ability writing under time pressure
- Exciting!
 - Finding has been replicated
 - Unusual in NLP

References

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

Laboratory ratings evaluation

- Probably most common type in NLG
 - Easiest and quickest
 - Well accepted in NLP literature
 - Less well accepted outside NLP
 - Don't try to publish this in a medical journal/conference!
- We know these may disagree with results of task performance eval!
 - Earlier Babytalk experiment, comparing text to visualisation
 - Doctors preferred visualisations
 - Doctors made better decisions from text

Human Evaluations

	Task	Ratings
Real-world	Most meaningful	intermediate
Laboratory	intermediate	Least meaningful

	Task	Ratings
Real-world	Most expensive	intermediate
Laboratory	intermediate	Cheapest

Metric-based evaluation

- Create a gold standard
 - Input data for NLG system (scenarios)
 - Desired output text (usually human-written)
 - Ideally 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

Metric evaluation: Experimental Design

- Hypothesis: System A texts more similar to gold standard texts than system B texts
- Subjects: none
- Material: Scenarios and gold standard texts
 - Ideally multiple gold standard texts for each scenario, since information can be communicated in several ways
- Procedure: Use scoring metric such as BLEU
- Statistics: specialised

Example: SumTime input data

Day/ Hour	Wind Direction	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

- Does similarity to gold-standard mean better for readers?
- Is SSW'LY better than SSW?
 - 2 out of 3 reference texts use SSW
 - Need 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 SumTime be evaluated against human texts?
 - SumTime texts are better than human texts!

Are metrics meaningful?

- Assess by validation study
 - Do “gold standard” evaluation of multiple NLG 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” evaluations?
 - Do any metrics correlate?

Validation Study: Experimental Design

- Hypothesis: BLEU, other metrics correlate with human ratings
- Subjects: 14 people who regularly read marine forecasts
 - Similar to SumTime subjects
- Material: Forecast texts produce by 6 marine forecast generators, plus the human forecast text, for 14 scenarios
- Procedure: Subjects rated texts on Clarity and Accuracy
 - Latin square design
- Analysis: Correlation (Pearson) between metrics and human ratings of the different system

Validation: result

- Clarity (readability): Best predicted by NIST-5 (BLEU variant)
 - Statistically significant correlation using “generous” stats
 - 1-tailed, no multiple hypothesis correction
 - Not significant under “conservative” stats
 - 2-tailed, multiple hypothesis correction
- Accuracy: Not predicted by any metric
 - No correlation is significant, no matter how calculated

References

- E Reiter and 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 lack confidence in metric-based evaluation of NLG systems
- I want to see validation studies that show
 - statistically significant correlation of 0.8 (or more)
 - Significant with conservative statistics
 - Against high-quality human evaluation
 - Real-world and/or task-based
 - With clarity about scope
 - Only valid when comparing statistical NLG systems?
 - Only valid in newswire genre?

Commercial Evaluation

Commercial Evaluation

- Contractual
- Cost
- Benefits
- Risks

Poorly understood!

Contractual

- Does the system meet the contractual requirements?
- Assessed by User Acceptance Test (UAT)
- Won't further discuss here

Cost

- Cost of system
 - Life cycle: Development, testing, maintenance
- Does statistical NLG reduce cost?
 - Claims that system can be developed (built) quicker using statistical NLG
 - Very hard to rigorously test such hypotheses
 - Lots of confounds: team skill, enthusiasm, background
 - What about testing and maintenance?
 - statistical NLG easier/ harder than rule-based?
 - “please change *rainy day* to *wet day* in the computer-generated forecasts”

Benefits

- Quantitative
 - Better task performance, reduced costs due to automation, etc.
- Qualitative
 - “Arria’s report generation system saves money, but even more important is that it improves consistency”
 - Very important, even if hard to measure
- Uptake
 - Will people use the system?
 - Medicine: very poor uptake of decision support
 - Change management

Risks

- Low-probability, high-impact events
 - Skillsum (assessment feedback) user starts crying when she sees the feedback
 - Poor baby care due to bad BT-Nurse advice?
- Undesirable side effects
 - “Adding localised weather forecasts to my website will help my users, but I am worried my Google search rank will go down”
- How do we measure/ evaluate these?

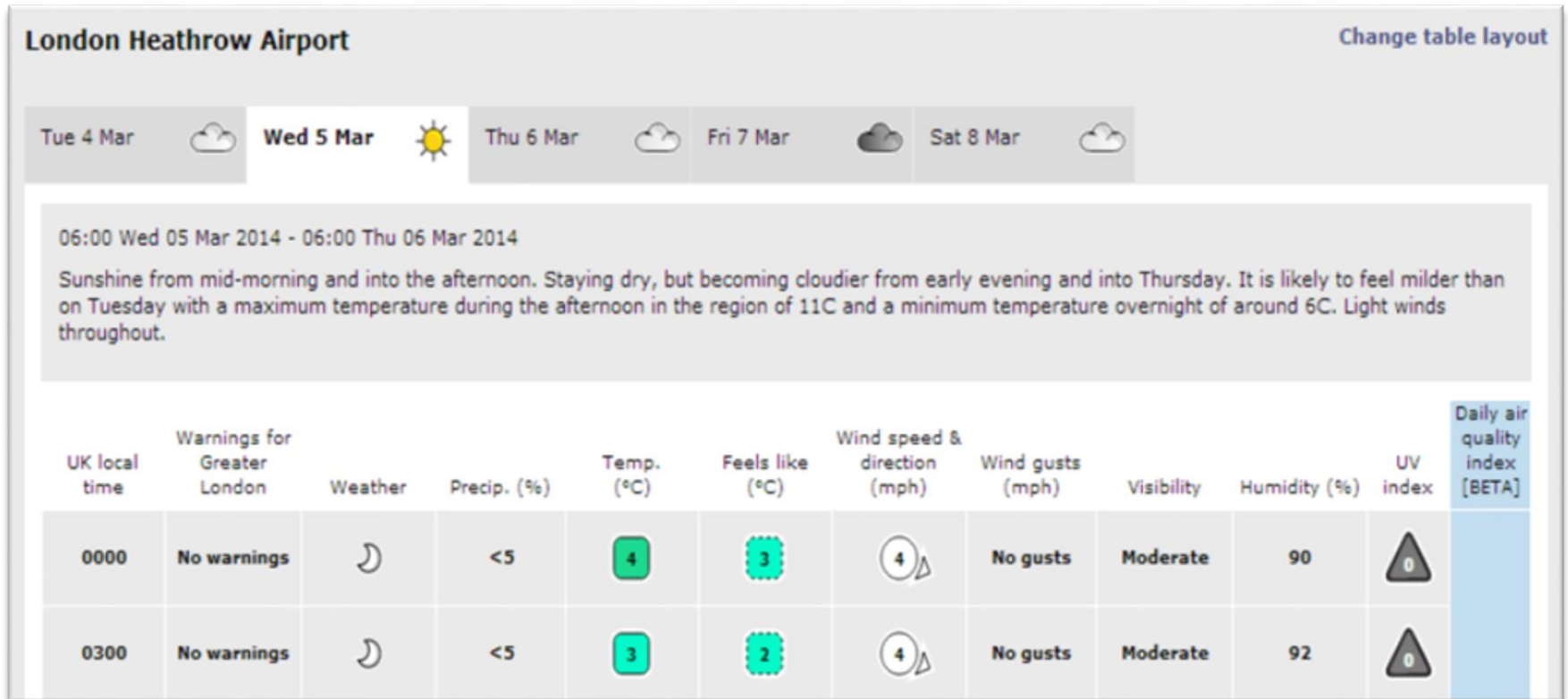
Worst-Case performance

- We need to understand the impact when a system performs really badly
- Not just performance on average
 - Focus of academic evaluations

Case Studies

Case Study 1: UK Public Forecasts

- Arria system which created public weather forecast texts.
- Supplement standard weather graphics



Evaluation

- System was deployed on UK Met Office “Invent” (beta software) web site
- Users could be asked to respond after reading a forecast.
- How can we evaluate the forecast generator in this context?

Experimental Design

- Hypotheses
- Subjects: *had to be genuine users who came across system. Not allowed to recruit subjects*
- Material
- Procedure
- Analysis

Actual experiment

- <http://www.aclweb.org/anthology/W14-4401>

Actual experiment

























- Hypotheses: Users value forecast texts as useful supplement to standard weather graphics
- Subjects: naturally occurring (35 people)
- Material: forecast and graphics, for any UK location
- Procedure: respond to three questions
 - helpfulness, length, would you recommend
- Analysis: Percentage of responses (no stats or p value)

Results

- 97% found the text to be helpful
- 74% thought length was about right
- 91% would recommend texts be included in local public forecasts

Case Study 2: Public Forecasts in Galicia

- Public forecasts for Galicia (including Santiago)

9th December, Monday			10th December, Tuesday			11th December, Wednesday			12th December, Thursday		
Morn.	Aft.	Night	Morn.	Aft.	Night	Morn.	Aft.	Night	Morn.	Aft.	Night
											
											
Min: 1° Max: 14°			Min: 5° Max: 16°			Min: 7° Max: 16°			Min: 11° Max: 15°		

There will be clear skies at the beginning and towards the middle of the term, although at the end they will be very cloudy. We expect precipitations on Thursday morning. The temperatures will be normal for the minimums and high for the maximums for this period of the year, with minimums in notable increase and maximums without changes.

Evaluation

- Ask forecasters to assess quality of forecasts

Experimental Design

- Hypotheses
- Subjects: *forecasters*
- Material
- Procedure
- Analysis

Actual experiment

- https://www.researchgate.net/publication/266317013_Linguistic_Descriptions_for_Automatic_Generation_of_Textual_Short-Term_Weather_Forecasts_on_Real_Prediction_Data

Actual experiment

- Hypotheses: Weather forecasters think forecasts are high quality
- Subjects: 1 forecaster
- Material: 45 datasets and computer-gen forecasts
- Procedure: respond to several questions
- Analysis: mean and standard deviation (no p value)

Results

Indicate in which degree you identify the type of results expressed as the type of results expressed by yourself:

Sky cover:	5	(5 is maximum value)
Precipitation:	5	
Wind:	5	
Temperature	5	

Do you agree with the provided descriptions:

Sky cover:	4.97
Precipitation:	4.53
Wind:	5
Temperature	5

Indicate in which degree the vocabulary is used correctly: 5

Indicate in which degree the content is correctly grouped to facilitate the comprehension of the description: 4.64

Indicate in which degree the format of the report, including the punctuation, is the most adequate: 4.53

Case Study 3: Error Analysis in Image Desc

- Automatic generation of descriptions/captions for images

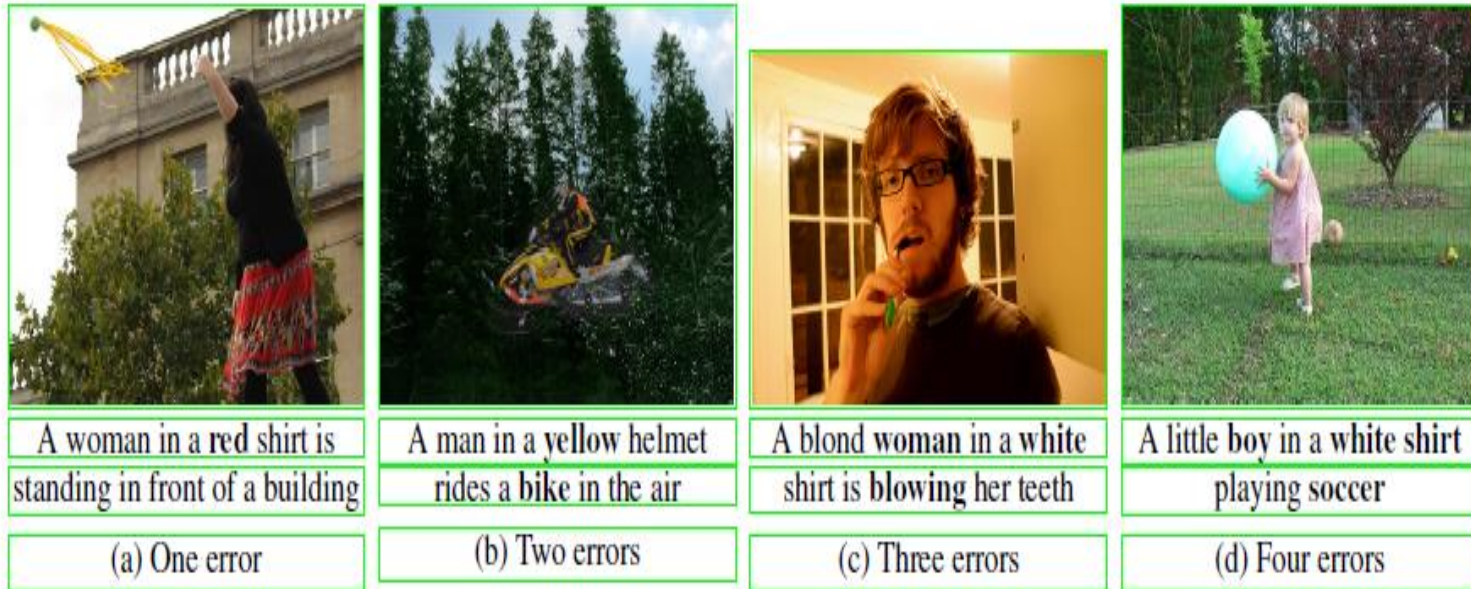


Figure 1: Examples of images with 1–4 errors. The annotated errors are marked in boldface.

Evaluation

- Analyse the type of errors made by an existing image description algorithm

Experimental Design

- Hypotheses: *none*
- Subjects
- Material
- Procedure
- Analysis

Actual experiment

- <https://arxiv.org/abs/1704.04198>

Actual experiment

- Hypotheses: none
- Subjects: 2 annotators
- Material: 1014 generated descriptions of Flickr images
- Procedure:
 - Develop annotation scheme for types of errors
 - Annotate descriptions under this scheme
- Analysis: annotation frequency; inter-annotator agreement

Results

Type	Count	Type	Count	Type	Count
generally unrelated	264	non-existent object	47	color	14
color of clothing	195	age	40	non-existent subject	11
activity	168	stance	38	wrong-object	7
type of clothing	104	position	37	similar-subject	3
gender	98	extra subject	34	extra object	1
scene/event/location	91	similar-object	31	wrong-subject	1
number	61	other	20		

Table 2: Number of times each error was annotated in our fine-grained analysis.

Interannotator agreement: 55% on precision/recall basis

Concluding Thoughts

Summary

- Many ways to evaluate NLG
 - Task-based (real-world or laboratory)
 - Human ratings (real-world or laboratory)
 - Metrics
 - Commercial (poorly understood)

Which technique to use?

- Most common is laboratory ratings
- Task-based and/or real-world evaluation is harder, but more meaningful.
- Metrics should not be only evaluation
- Good experimental design and statistics!
 - Getting the basics right is most important thing!

Challenges

- Education and spread of best practice
- Commercial evaluation
 - How do we measure costs, benefits, risks?
- Design cheap/quick human evaluation that correlate with high-quality human evaluation
- Well-validated metrics

Personal Lessons

- Publish negative results
 - Edge cases and outliers matter
 - Make sure code is debugged
 - Validate metrics
 - Look at risks and worst-case performance
-
- Not rocket science
 - But very important

References

- I have several blogs about evaluation on ehudreiter.com

Evaluation in NLG/NLP

- Lets do it properly and rigorously, with replicable results!
- Otherwise its not science