Information Retrieval (IN4325)

NLP Evaluation

Dr. Nava Tintarev

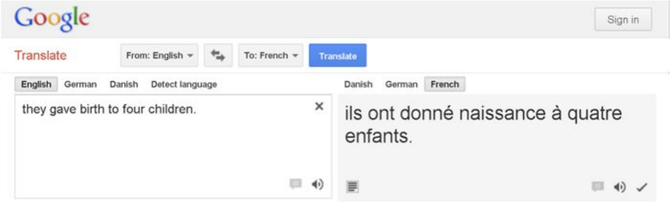
Assistant Professor, TU Delft

Credits: These slides are modified from Prof. Ehud Reiter
(Abdn Uni and Arria/Data2text)

P8: The social impact of natural language processing

Biased measured by Implicit Association Test in language models

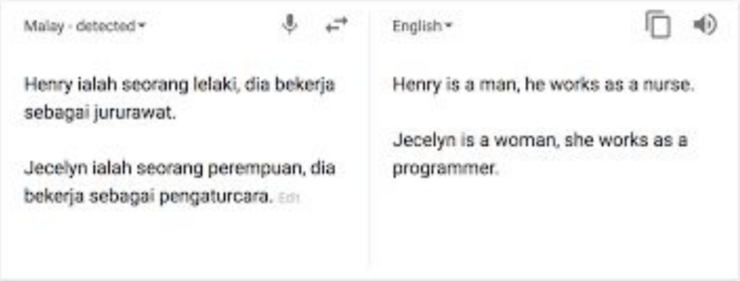
[Caliskan-Islam et al, 2016]



they gave birth to four children.

ils ont donné naissance à quatre enfants.

As the screenshot above shows, Google Translate uses the masculine plural ("ils") for the English "they," even in cases where context indicates that the feminine plural ("elles") is intended.

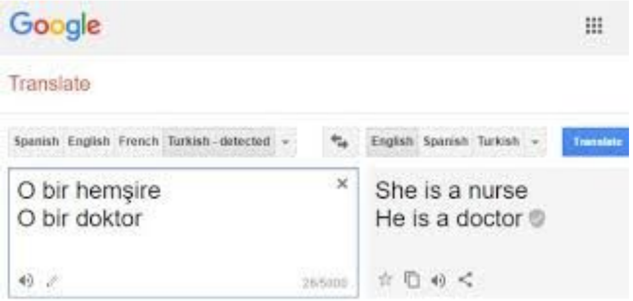


Henry ialah seorang lelaki, dia bekerja sebagai jururawat.

Jecelyn ialah seorang perempuan, dia bekerja sebagai pengaturcara.

Henry is a man, he works as a nurse.

Jecelyn is a woman, she works as a programmer.

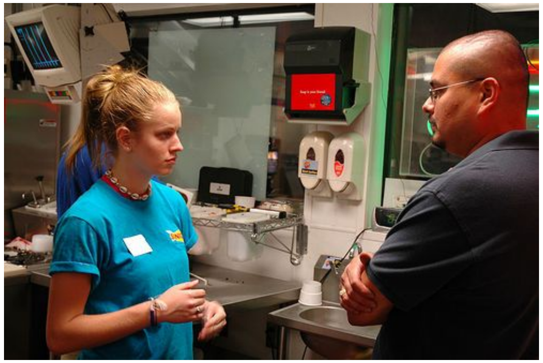


O bir hemşire
O bir doktor

She is a nurse
He is a doctor

Stereotypical descriptions in Flickr30K dataset
[van Miltenburg, 2016]

1. A blond girl and a bald man with his arms crossed are standing inside looking at each other.
2. A **worker** is **being scolded** by her **boss** in a **stern lecture**.
3. A **manager** talks to an **employee about job performance**.
4. A hot, blond girl **getting criticized by her boss**.
5. Sonic employees **talking about work**.



P8: The social impact of natural language processing

Exclusion

Overgeneralization

Topic overexposure

Bias confirmation

Detecting personal characteristics

False positives

Dual-use

Some solutions

exclusion: downsampling or priors;

overgeneralization: (explanatory) dummy variables, regularization, error weighting, or confidence thresholds

General caution when applying and interpreting results!

Last week

- Semantics
- Word sense
- Wordnet
- Path based similarity
- Information content similarity
- Lexical choice
- WSD

This week

- Evaluation
- Natural Language Generation
- Task (extrinsic) evaluation
- Human ratings (intrinsic) evaluation
- Metric evaluation
- Setting up statistical tests
- Concluding thoughts

TU Delft

Evaluation

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

Types of NLP Evaluation

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

- Controlled vs Real-World

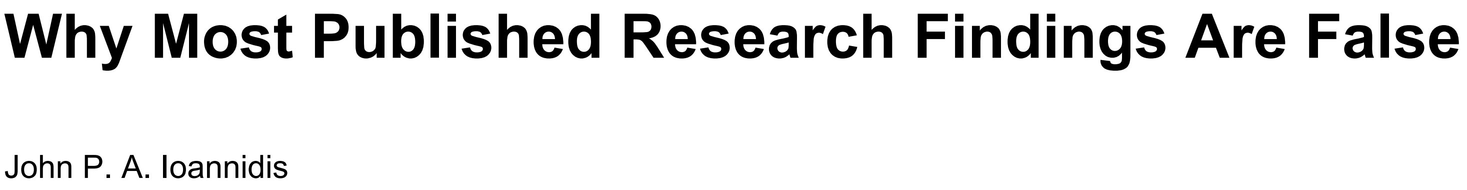
Task-Performance Eval

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

Aside: Hypothesis Test Angst

- Angst/ doubts/ concerns about failure to replicate “significant” findings
  - 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 NLP findings?



Why Most Published Research Findings Are False

John P. A. Ioannidis

Warning signs:

- 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

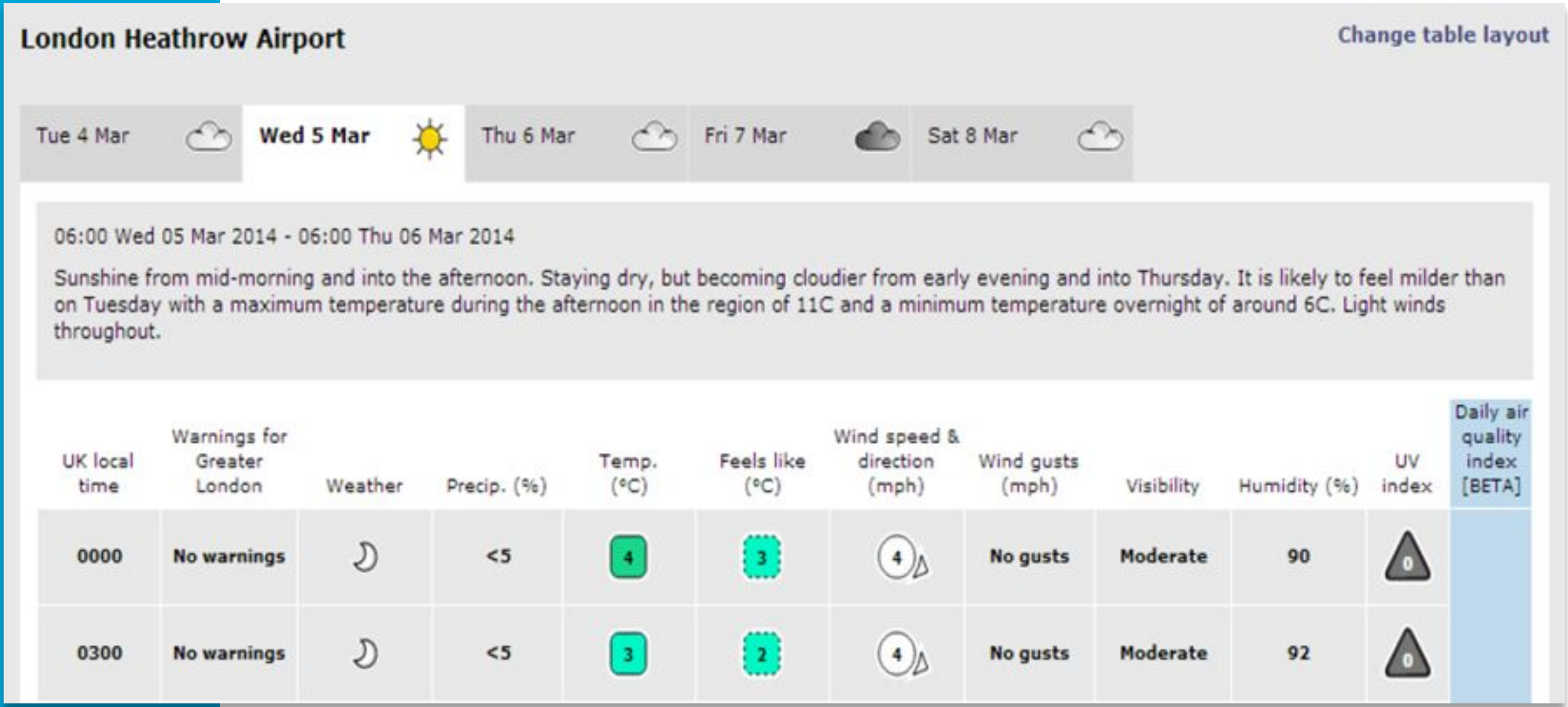
Natural Language Generation

COMFREAK@PIXABAY

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.

Natural Language Generation



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

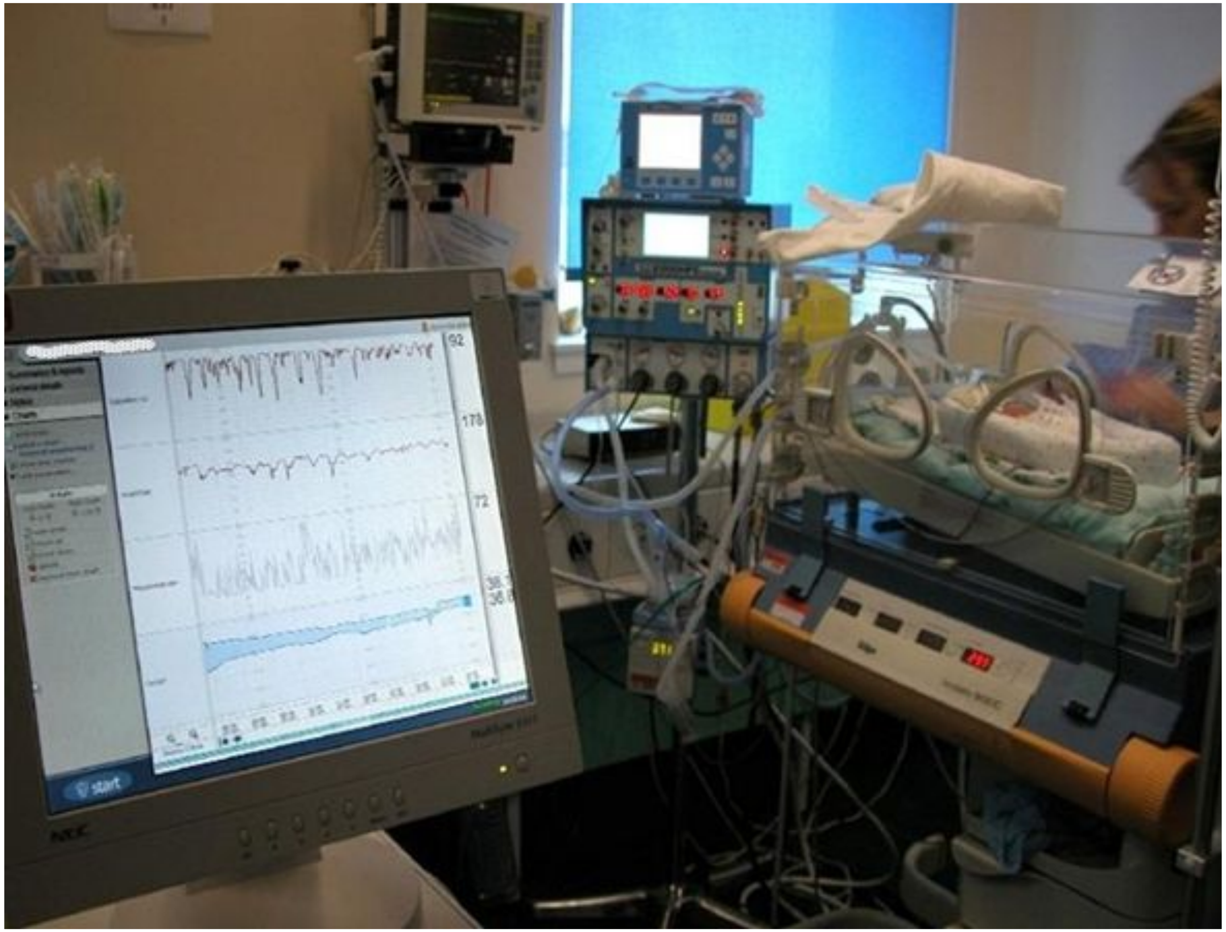
Sunshine from mid-morning and into the afternoon. Staying dry, but becoming cloudier from early evening and into Thursday. It is likely to feel milder than on Tuesday with a maximum temperature during the afternoon in the region of 11C and a minimum temperature overnight of around 6C. Light winds throughout.

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	

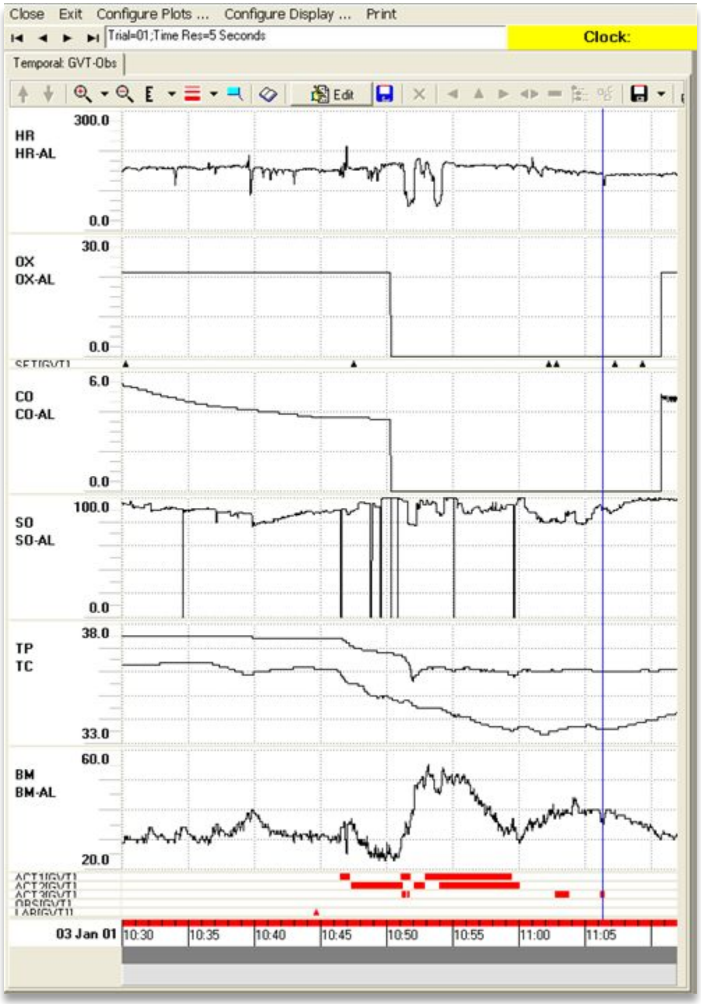
Complex: BabyTalk

- 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


Trial=01;Time Res=5 Seconds

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 (FiO2) 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

Real World or Controlled?

- **Real world evaluation** (doctors want this)
  - Install in ward, used by medics caring for babies
  - *Ecological validity*
- **Controlled evaluation** (psychologists want this)
  - Compare Babytalk against a control
  - Minimise differences in scenarios, subjects, etc
  - Easiest done in (artificial) *laboratory setting*
- If can not have both, which do we choose?

Task (Extrinsic) Evaluation

Real-world and laboratory

NLG Evaluations

	Task (extrinsic)	Ratings (intrinsic)	Metric
Real-world	Stop	BT Nurse	NA
Laboratory	BT 45	Sumtime	*weather*

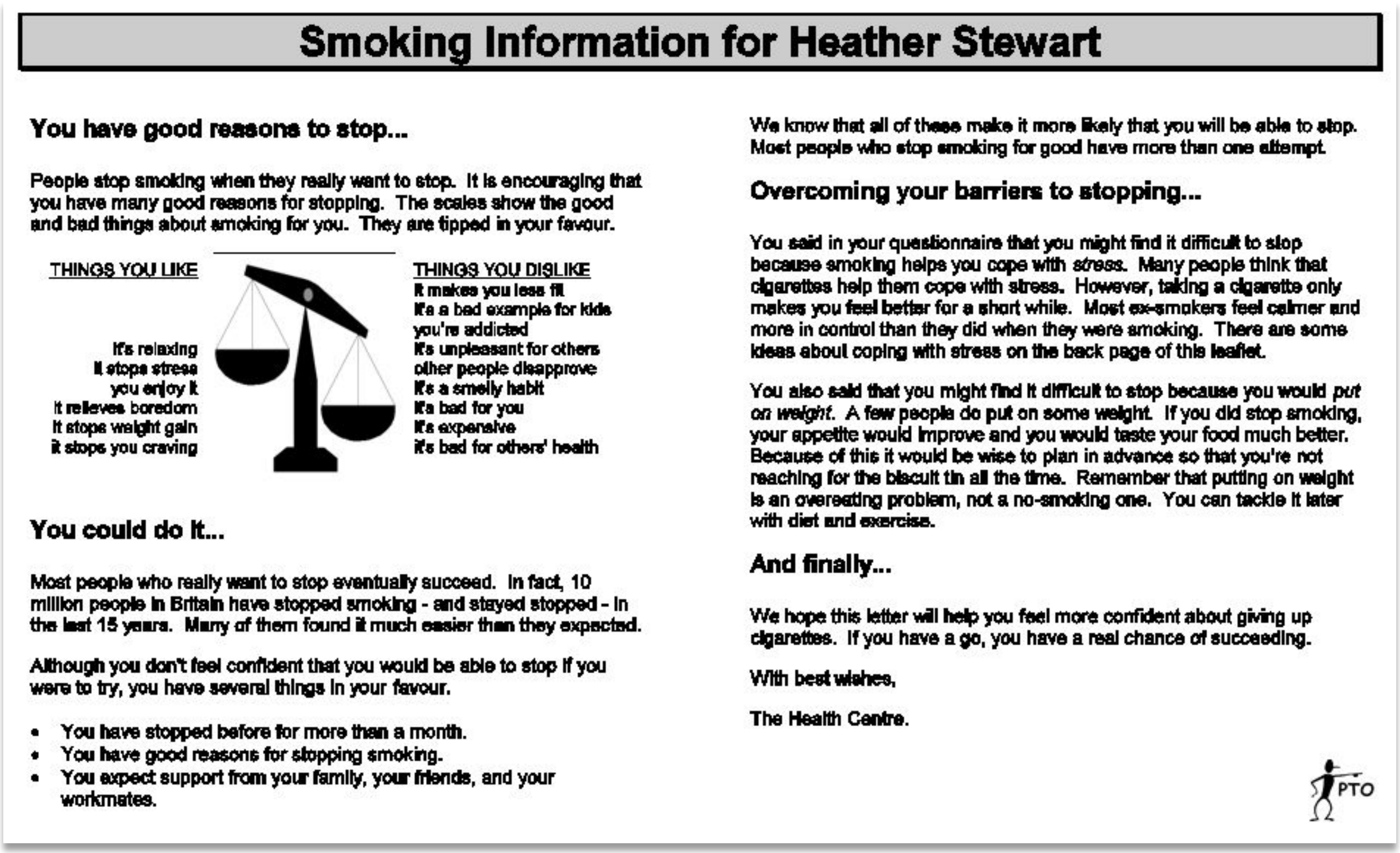
Task-Performance Evaluation

- Extrinsic evaluation
- Measure whether NLG system achieves its goal
  - Better decision making
  - Better clinical outcome
  - Etc.
- Evaluate in real world or in laboratory 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 measured smoking cessation rates in the groups

STOP example



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.

PTO

Real world: STOP smoking

- 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

- 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
- NLP needs more negative results

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

Laboratory experiment: BT45

- Babytalk BT-45 (decision support)
- Chose 24 data sets (scenarios)
  - From historical data (5 years old)
- Created 3 presentations of each scenario
  - BT45 text, Human text, Visualisation
- Asked 35 clinicians to look at presentations and choose intervention (in 3 min)
  - In experiment room, not in ward!
  - Compared intervention to gold standard

Results: BT45

- Correct decision made
    - BT45 text: 34%
    - Human text: 39%
    - Visualisation: 33%
- Note:
    - BT45 texts mostly as good as human, but did poorly when desired intervention was “no action” or “reattach sensors”.

Edge cases matter

System needs to perform well in atypical as well as normal cases:

- “reattach sensor” as well as “increase oxygen level”
- In NICU, “no action” is probably most common decision, needs to be handled well!

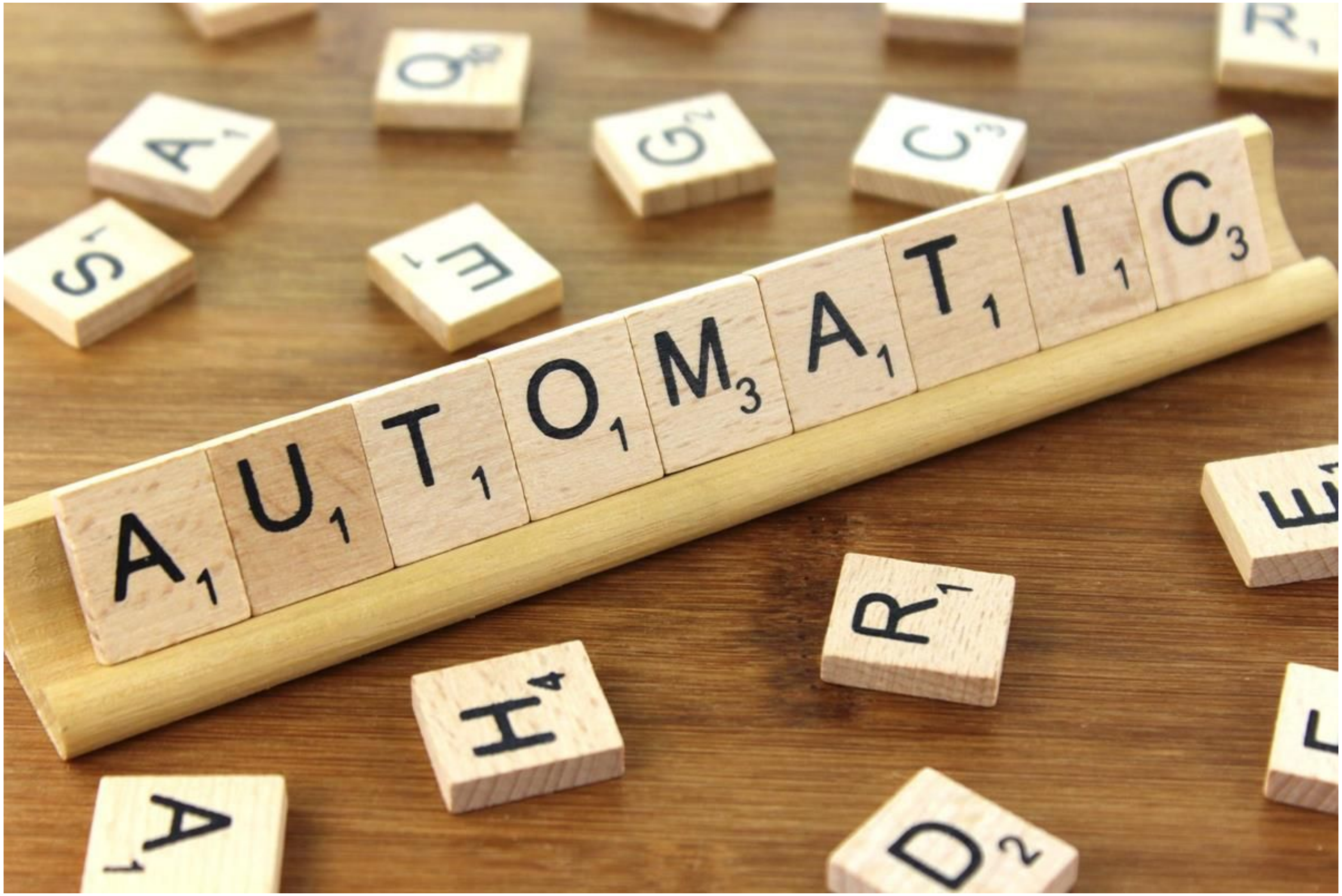
Reference

- 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
- Evaluation is of specific system, not generic algorithm or idea
  - Small changes to BT45 (STOP?) would have significantly changed evaluation result

Human Ratings (Intrinsic) Evaluation
Real-world and laboratory



Human Ratings

- Ask human subjects to assess texts
  - Readability (linguistic quality)
  - Accuracy (content quality)
  - Usefulness
- Intrinsic evaluation
- Usually use Likert scale
  - Strongly agree, agree, undecided, disagree, strongly disagree (5 pt. scale)

Real world: BT-Nurse (expert eval)

- Deployed BT-Nurse on ward
- Nurses used it on real patients
  - Vetted to remove texts which could damage care
- Nurses gave scores (3-pt. scale) on each text
  - Understandable, accurate, helpful
  - Agree, neutral, disagree
- Also free-text comments
- No baseline/control

Results: BT-Nurse

- **Numerical results**
  - 90% of texts understandable
  - 70% of texts accurate
  - 60% of texts helpful
  - [no texts rejected as potentially harmful]
- **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

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
- **Arria** (as a commercial company) puts a lot of effort into fixing bugs and quality assurance more generally

Laboratory experiment: Sumtime

- Marine weather forecasts
- Chose 5 weather data sets (scenarios)
- Created 3 presentations of each scenario
  - Sumtime text
  - Human texts (actual forecaster text)
  - 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

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.

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

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

Human ratings evaluation

- Probably most common type in NLG
  - Well accepted in NLP literature
  - Less well accepted outside NLP
- Easier/ quicker than task-based
  - For laboratory evaluation, can *sometimes* use crowd computing (e.g., Mechanical Turk)
    - Not always, and sometimes MTurk studies need to be rerun outwith MTurk
  - Can answer questions which are hard to fit into a task-based evaluation
    - Can ask people to generalise

Human Evaluations

	Task (extrinsic)	Ratings (intrinsic)
Real-world	**Most meaningful**	intermediate
Laboratory	intermediate	*Least meaningful*

	Task (extrinsic)	Ratings (intrinsic)
Real-world	**Most expensive**	intermediate
Laboratory	intermediate	*Cheapest*

Questions?

Metric evaluation

Metric-based evaluation

- Create a gold standard
  - 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, ROUGE, METEOR**
- Widely used in machine translation

Metrics

Summary scored by amount of N-gram overlap between candidate and human-generate summary.

BLEU

- Average number of overlaps of different length.

ROUGE - Recall-oriented Understudy for Gisting Evaluation

- Length of n-gram is fixed (ROUGE-1, ROUGE-2, ROUGE-n)
- ROUGE-L: longest common subsequence
- ROUGE-S/SU: number of skip bigrams (U+unigrams)

- A skip bigram is a pair of words in their sentence order, but allowing for any number of other words to appear between the pair.

Metrics

BLEU

- Machine translation
- how much n-grams in *machine generated translations* appear in human reference translations.
- **Precision** based measure

ROUGE - Recall-oriented Understudy for Gisting Evaluation

- Summarization, but not as useful as BLEU in this task
  - Humans are inconsistent!
- how much n-grams in the *human reference summaries* appear in machine generated summaries
- **Recall** based measure

Metrics

$$F_{mean} = \frac{10PR}{R + 9P}$$

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**
  - Unigram. Harmonic mean of precision and recall
  - Recall weighted higher than precision
  - Stemming and synonym matching
  - Seems to work better than BLEU compared to human judgement on sentence/segment level (0.964 versus 0.403)
  - BLEU should be better on corpus level (0.817)

- **Edit distance**
  - minimum number of operations required to transform one string into the other
  - Insertion, deletion, substitution.

- There are many others: NIST (weighted BLEU), distinct...
- And of course classical IR measures: Precision, Recall, F-scores

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

- 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!

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

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

- Validation study in weather domain
  - Gold standard was human (reader) ratings of readability and accuracy (not usefulness)
    - Not ideal as “gold standard” evaluation
- Readability: Best predicted by NIST-5 (BLEU variant)
  - Decent correlation for similar systems
  - Less good for different sys
- Accuracy: Not predicted by any metric

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

- Many limitations
  - we don’t have strong evidence that metrics predict human ratings, let alone task performance
  - 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

- **Alternative: Pyramid Method** (Nenkova et al, 2007)
  - Human labels to compare candidate and reference
  - Shared Contents of meaning, ranked by importance

Metric-based evaluation

- I lack confidence in metric-based evaluation of NLG systems
  - I want to see correlation of 0.8 (or more) with high-quality human evaluation
  - Clarity about scope (e.g. does metric only work when comparing statistical NLG systems?)

- Strength of validation evidence for metrics in other areas of NLP?

Set up hypotheses and stats

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

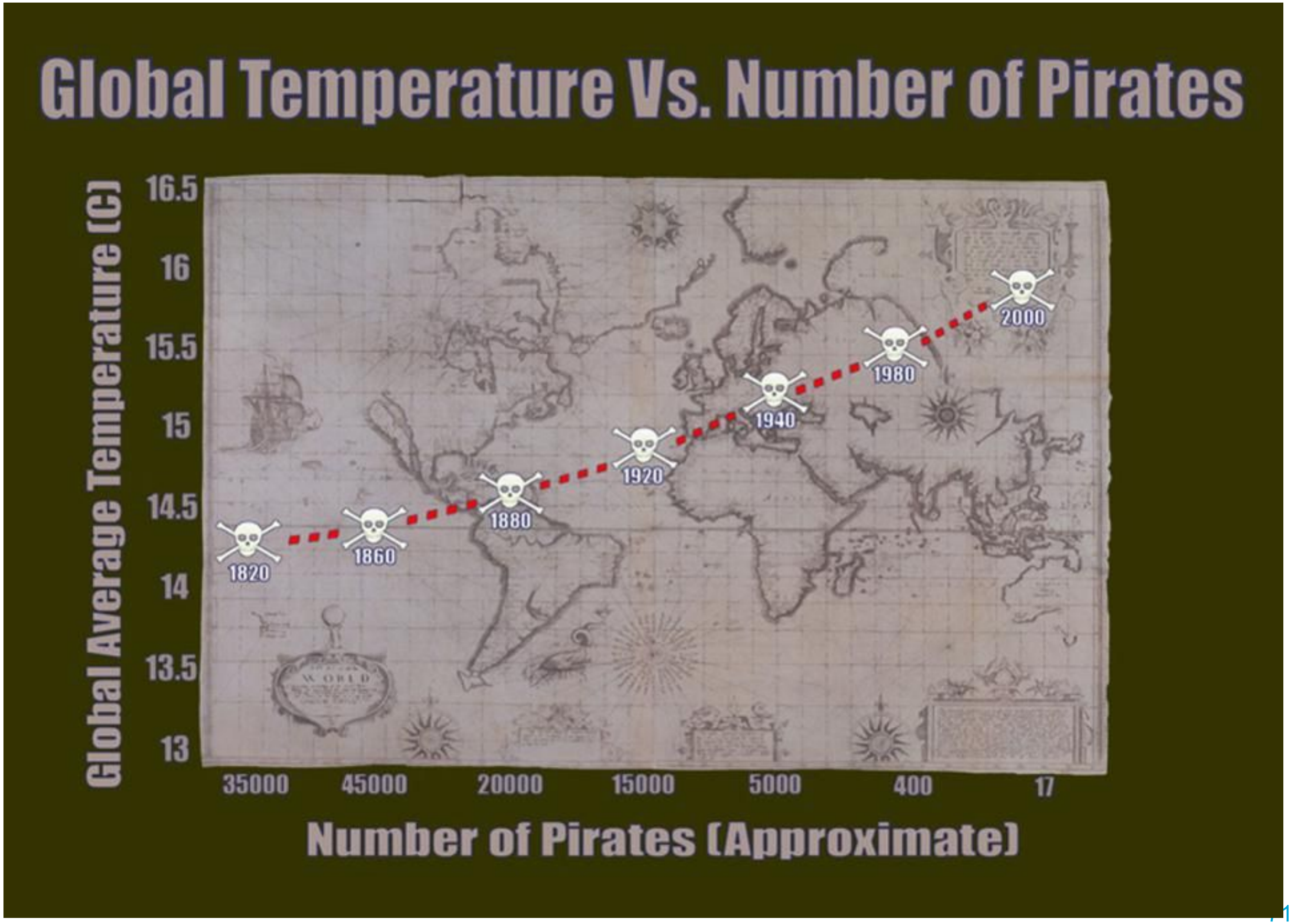
- Hypotheses
- Subjects
- Material
- Procedure
- Analysis

Hypotheses: before experiment

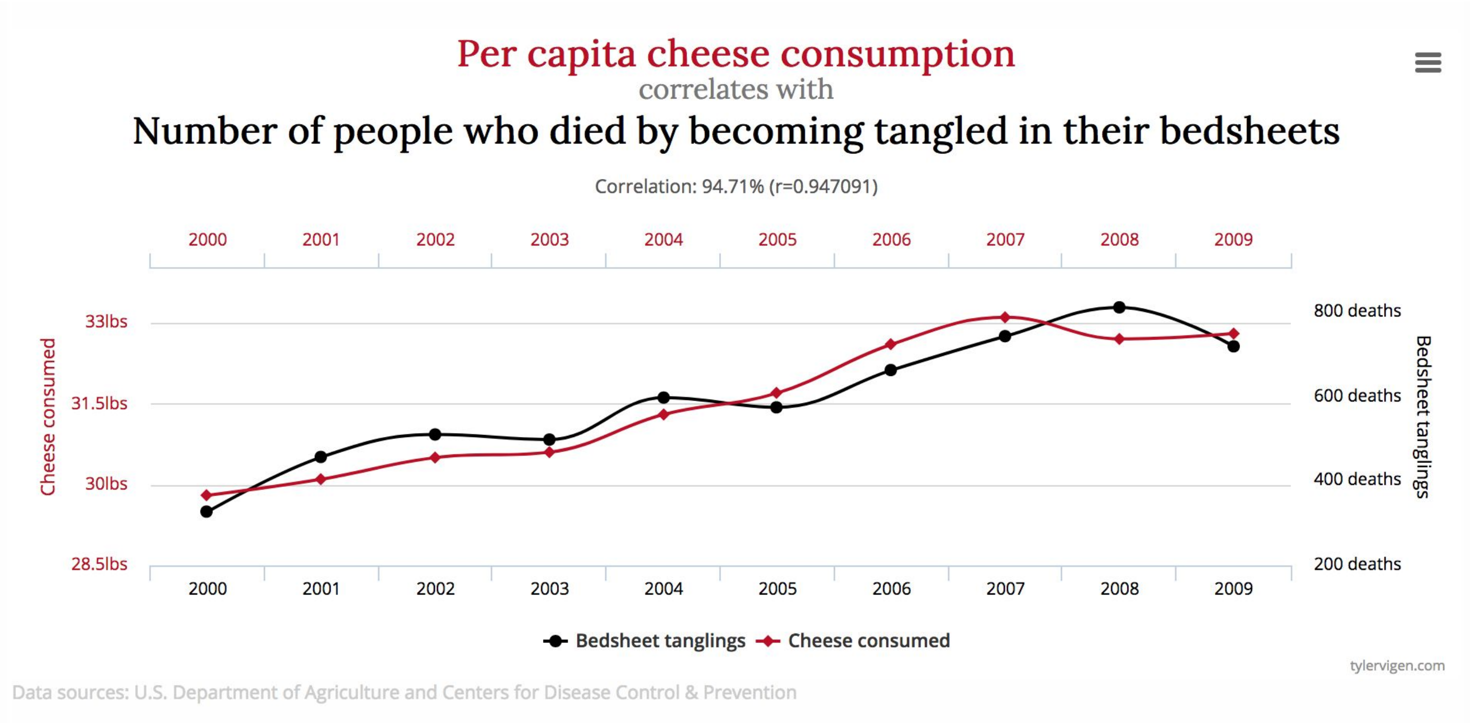
- Define hypotheses, stats, etc in detail before the experiment is done
  - » In medicine, expected to publish full experimental design beforehand
    - – https://clinicaltrials.gov/
  - » More generally, Open Science Frameowrk (OSF). Example: https://osf.io/65z4h/
  - » If multiple hypothesis, reduce p value for significance (discuss later)
- Why?

Correlation and causation

More: http://www.tylervigen.com/spurious-correlations



Correlation and causation



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

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
  - » E.g., 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
  - » E.g., 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
  - » Depends on expected effect size
    - – More subjects needed for smaller effects
  - » Typically looking for 50+
    - – Not a problem with human computation
    - – Can be real hassle if need subjects with specialised skills or backgrounds
  - » Effect size calculators

Recruitment of subjects

- General subjects (easier)
  - » Human computation (Figure8, MTurk…)
  - » 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 a certain date
  - Wanted human/control texts from each of 5
- **Number of presentations:** 3
  - SumTime (main)
  - Human (control)
  - Hybrid (of content-det vs microplan/real)
- **Procedure:**
  - 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
- Consent forms and ethics committee!

Procedure: Exclusion

- When do we drop a subject from the experiment?
  - » Incomplete responses?
  - » Inconsistent responses?
  - » Bizarre responses?
- Human-computation
  - » Acceptance rates
  - » Control questions
  - » Durations

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!
    - Need to check if data is normally distributed.

Statistics: Normalisation

- Some users are more generous than others
- Some scenarios are harder than others
- Potential bias
  - User X always rates “Great”, Y always “Poor”
  - 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, does not matter how generous he is (“great” vs “poor”)
- Multiple hypotheses: $p < 0.025$
  - » Because 2 hypotheses

Which technique to use?

- Most common is laboratory ratings
  - But we know these may not correlate with task performance (e.g. Babytalk experiment)
- Task-based and/or real-world evaluation is harder, but more meaningful.
- Metrics should not be only evaluation
- Good experimental design and statistics!

Challenges

- Education and spread of best practice
- 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

- Not rocket science….still very important!
- Do it properly and rigorously, with replicable results!

- Otherwise it is not science

This week

There are a lot of ways to evaluate...

- Task (extrinsic) evaluation
- Human ratings (intrinsic) evaluation
- Metric evaluation
- Setting up statistical tests
- Concluding thoughts

Next week

Machine learning for NLP

- Classes of machine learning problems
- Feature selection/extraction
- Common ML techniques
  - Discriminative: SVM, MaxEnt
  - Generative: NB, logistic regression
  - Discriminative v. Generative
- Application domains
  - NER
  - Fake review detection

Next milestones

- Review P10: handed out March 22, due March 29.
- NLP intermediate project report: **due April 5 (in 1 week)**

Questions?