NLP models as vaccines for language problems

Significant lessons from experimental sciences

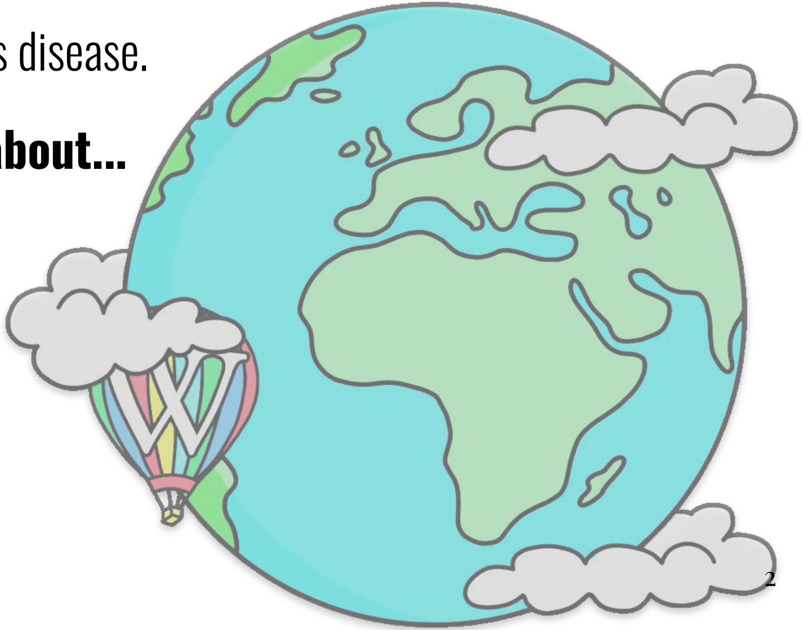
Year 3000...

The Earth is finally a safe and pleasant place for humans again.

However, 1000 years of global warming released a **dangerous bacteria from the permafrost**.

The bacteria starts to infect human hosts, causing a mysterious disease.

Centuries in insipid watery ice made the bacteria **obsessive about...**



...vanilla ice-cream!

The illness is called:

Compulsive
Obsessive
Vanilla
Ice-cream
Disease

Chaos!

The bacteria spreads rapidly, and infected humans start eating **tons of vanilla ice-cream**.

Milk prices rise to the stratosphere, ice-cream makers strike, diabetes and obesity break records...

Governments impose ice-cream lockdowns, interplanetary travel is forbidden, panic everywhere!

After months of an unprecedented crisis,

a lab finally announces a vaccine at phase 3!

In phase 3, a vaccine is evaluated using an experiment called **randomized control trial**

Randomized control trial

Group A

Vaccine

Group B (control)

Placebo

Conclusion:

The vaccine works.

What a relief for humanity!

After 1 month, average nb. of ice-creams/day (ICD):

- Group A: ICD_A=1.47
- Group B: ICD_B=1.56

But... maybe humans forgot all about statistics?

- Is the difference/effect observed in this experiment significant?
  - ICD_A=1.47 ice/creams per day
  - ICD_B=1.56 ice/creams per day
  - **δ=0.9**
- Maybe the **sample** is too small or biased to conclude that vaccine (A) is better than placebo (B).

Given the samples, the metrics, and the experiment's conditions:

What is the probability of making a false claim when assuming A ≠ B **in general**?

⇒ p-value!

What about NLP?

- Group A and group B could be two NLP **models/systems** we want to compare
  - A = our system, B = baseline or state-of-the-art system
  - **Is system A really better than system B?** (is vaccine (A) really better than placebo (B)?)
- Empirical/experimental science has become the norm in NLP
- We do not care (enough) about conclusions drawn from our experiments
  - We do not systematically test for the statistical significance of our results
  - When we do it, we do not always apply the tests correctly
  - Our samples/test sets and measures are full of biases
  - Our experiments are not reproducible nor replicable
- NLP lacks rigorous **methodological standards** for reporting experimental results

Experiments in NLP

- Group A and group B are two **models/systems** we want to compare
  - Is A better than B?
- $x = x_1 \ldots x_n$ is a **test set** composed of n items
- A is applied to X to get the **evaluation measure** M, and same for B
  - Suppose M(A,X) > M(B,X) → the higher the better, so it seems like A >> B
- The observed **difference** (or effect) is $\delta_{A\text{-}B}(x) = M(A,x) - M(B,x)$

Can this difference be due to chance?

Would we observe a similar value for a new independent test set x'?

How likely is the observed outcome if A was no better than B in general?

Hypothesis testing

- We formulate this as a hypothesis test*:
  - H_0: $\delta(X) \leq 0$ $\Rightarrow$ if this is true, the observed difference is not significant, so A is no better than B
  - H_1: $\delta(X) > 0$
- If we reject H_0 $\Rightarrow$ the difference is significant (>0)
- The **p-value** is the probability of observing the difference $\delta_{A\text{-}B}$ under the null hypothesis, that is:
  - $P(\delta(X) \geq \delta_{A\text{-}B} | H_0)$ $\Rightarrow$ probability of rejecting H_0 when it is actually true

* Random variable X represents all possible n-sized test sets

Type I and type II errors

- Type I error: false positives
  - Rejecting H_0 when it is actually true, OR
  - Concluding that the observed difference greater than 0 ($A >> B$) but it actually isn't ($A \leq\leq B$)
  - If p-value is below the **significance** level (usually α=0.05), we say that the difference is **statistically significant**
  - In other words, if probability of making type I errors (p-value) is sufficiently low, we can reject H_0

- Type II error: false negatives
  - Not rejecting H_0 when it is actually false
  - Concluding that the observed difference is no greater than 0 ($A \leq\leq B$) but it actually is ($A >> B$)
  - A **test's power** is its probability of avoiding type II errors

- Goal :
  - Guarantee that the probability of type-I errors is upper bounded by α
  - Achieve as high power as possible

Difference of means

- Remember the average number of ice-creams/day:
  - ICD_A=1.47 ice/creams per day
  - ICD_B=1.56 ice/creams per day
  - Suppose also that
    - groups A and B have n=25 subjects
    - standard error of the difference is se=0.08

- Averages are normally distributed (remember the central limit theorem)
- Subjects are independent and identically distributed (iid) in groups A and B

- ⇒ Paired Student's **t-test** for the difference of means

  $T = ICD_A - ICD_B / (se / \sqrt{n}) = 5.625$ → **test statistic** (lookup p-value in table, n-1 degrees of freedom)

- In practice, e.g. scipy's stats.ttest_rel

More accurate tests (Yeh 2000)

- Precision (TP/P), recall (TP/T) and F-measure (2PR/(P+R))
- Recall has a simple formula, linearly dependent on TP
  - T is a constant of the test set x
  - We could use a paired t-test
- Precision and F-measure have more complex forms
  - Use randomized permutation test (Noreen 1989)

Randomized permutation (Noreen 1989)

Input: test set $x=x_1...x_n$, predictions $A(x_i)$ and $B(x_i)$ for systems A and B for each item x_i, measure M

1. Calculate the observed difference $\delta_{A-B}(x) = M(A,x) - M(B,x)$
2. Repeat R times (R is of the order of 10k to 100k)
3. For each item x_i in x
4. Exchange predictions $A(x_i)$ and $B(x_i)$ with probability ½
5. If the difference on the scrambled dataset is larger than $\delta_{A-B}(x)$
6. $r = r+1$
7. Return estimated p-value = $(r+1)/(R+1)$

Bootstrap (Efron & Tibshirani 1993)

Input: test set $x=x_1...x_n$, predictions $A(x_i)$ and $B(x_i)$ for systems A and B for each item x_i, measure M

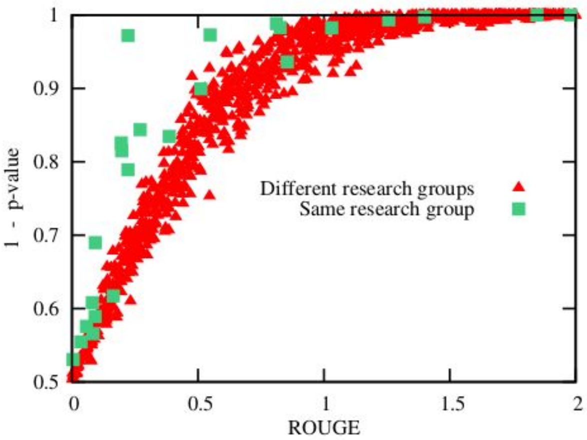
1. Calculate the observed difference $\delta_{A\text{-}B}(x) = M(A,x) - M(B,x)$
2. Repeat R times (R is of the order of 10k to 100k)
3. Randomly draw a new a new n-sized test set x' from x with replacement
4. Calculate the difference $\delta_{A\text{-}B}(x')$ on the new test set
5. If $\delta_{A\text{-}B}(x') > 2\delta_{A\text{-}B}(x)$
6. $r = r+1$
7. Return estimated p-value = $(r+1)/(R+1)$

Practical considerations for sampling-based tests

- Pre-calculate number of true positives, trues and positives for each test set item
- Permutation test
    - Only exchange items with differences, the test remains constant
- Bootstrap assumes sample distribution = population distribution (selection biases?)

Empirical investigation (Berg-Kirkpatrick et al. 2012)

- Relation between observed difference and p-value
- Summarisation (ROUGE), parsing (UAS), translation (BLEU)
- Model types, test set size, domains
- "simple thresholds are not a replacement for significance tests"



What's in a p-value? (Søgaard et al. 2014)

- Selection and measure biases

	TA (b)	UA (b)	SA (b)	SA(w)
Bio	0.3445	0.0430	0.3788	0.9270
Chem	0.3569	0.2566	0.4515	0.9941
Spoken	<0.001	<0.001	<0.001	<0.001
Answers	<0.001	0.0143	<0.001	<0.001
Emails	0.2020	<0.001	0.1622	0.0324
Newsgrs	0.3965	0.0210	0.1238	0.6602
Reviews	0.0020	0.0543	0.0585	0.0562
Weblogs	0.2480	0.0024	0.2435	0.9390
WSJ	0.4497	0.0024	0.2435	0.9390
Twitter	0.4497	0.0924	0.1111	0.7853

Table 2: POS tagging p-values across tagging accuracy (TA), accuracy for unseen words (UA) and sentence-level accuracy (SA) with bootstrap (b) and Wilcoxon (w) ($p < 0.05$ gray-shaded).

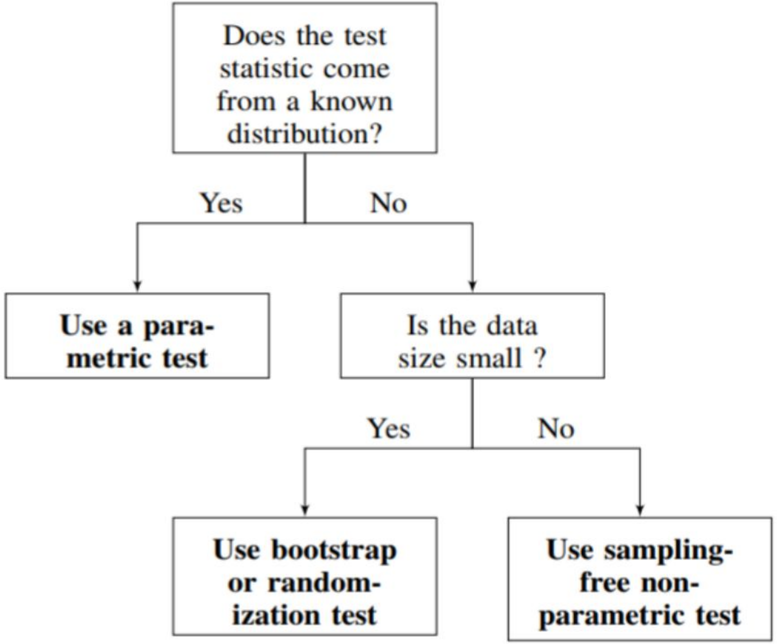
- Take-home message: report significance on all datasets and all metrics

What is the distribution of the evaluation measure?

- Parametric tests (known distribution)
  - Paired Student's t-test
- Non-parametric tests (unknown distribution)
  - Sampling-free (less powerful)
    - Sign test
    - McNemar's test
    - Wilcoxon signed rank test
  - Sampling-based (computationally expensive)
    - Permutation test
    - Bootstrap test

How to choose the right test?

Hitchhiker's guide (Dror et al. 2018)



Source: Dror et al (2018) *The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing*

Are we doing that in our papers?

General Statistics	ACL '17	TACL '17
Total number of papers	196	37
# relevant (experimental) papers	180	33
# different tasks	36	15
# different evaluation measures	24	19
Average number of measures per paper	2.34	2.1
# papers that **do not** report significance	117	15
# papers that report significance	63	18
# papers that report significance but use the **wrong** statistical test	6	0
# papers that report significance but do not mention the test name	21	3
# papers that have to report replicability	110	19
# papers that report replicability	3	4
# papers that perform cross validation	23	5

Source: Dror et al. 2018

Note: misuse of the word **significant**

Replication and reproduction

- Replication
  - Same models, different datasets
- Reproduction (Belz et al 2021)
  - Same models, same datasets
    - Same implementations
    - Different implementations

Multiple datasets, replicability (Dror et al. 2017)

- Multiple comparisons : probability of false claims increases
- Bonferroni's correction
  - Divide significance level α by the number of datasets N

Standard splits (Gorman & Bedrick 2019)

- We need to talk about standard splits
- Cross-validation for POS tagging
    - Significance test across splits
    - Bonferroni correction
- Some differences in standard test sets are not observed in cross validation
- Conclusions are not the same on two difference test sets

Show your work (Dodge et al. 2019)

- Influence of hyperparameters on results
- Adopted in EMNLP 2020's review forms and onwards

Impact of our conclusions and ethics

- Energy and policy considerations (Strubell et al. 2019)
- Stochastic parrots (Bender et al. 2021)
- State and fate of linguistic diversity (Joshi et al. 2020)
- Decolonising NLP (Bird 2020)
- ...

Take-home message

We need

Careful experimental design
Systematic significance tests
Frameworks for replicability
Awareness of biases in test sets
Avoid making false claims

We should improve methodological practices so that they may become standards in NLP one day...

Th$\alpha_{=0.05}$nks

Further reading on significance

Noreen 1989 *Computer intensive methods for testing hypotheses*

Efron & Tibshirani 1993 *An introduction to the bootstrap*

Yeh 2000 *More accurate tests for the statistical significance of result differences*

Berg-Kirkpatrick et al. 2012 *An Empirical Investigation of Statistical Significance in NLP*

Søgaard et al. 2014 *What's in a p-value in NLP?*

Dror et al. 2019 *Replicability Analysis for Natural Language Processing: Testing Significance with Multiple Datasets*

Dror et al. 2018 *The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing*

Further reading on reproducibility, diversity, ethics…

Strubell et al. 2019 *Energy and Policy Considerations for Deep Learning in NLP*

Dodge et al. 2019 *Show Your Work: Improved Reporting of Experimental Results.*

Korman & Bedrick 2019 *We Need to Talk about Standard Splits*

Joshi et al. 2020 *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*

Bird 2020 *Decolonising Speech and Language Technology*

Bender et al. 2021 *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*

Belz et al. 2021 *A Systematic Review of Reproducibility Research in Natural Language Processing*