TOWARDS REPRODUCIBLE ML RESEARCH IN NLP

Ana Lucic, Maurits Bleeker, Samarth Bhargav, Jessica Zosa Forde, Koustuv Sinha, Jesse Dodge, Sasha Luccioni, Robert Stojnic

ACL 2022

REPRODUCIBILITY IN NLP

Jesse Dodge, Sasha Luccioni, Jessica Zosa Forde

OVERVIEW

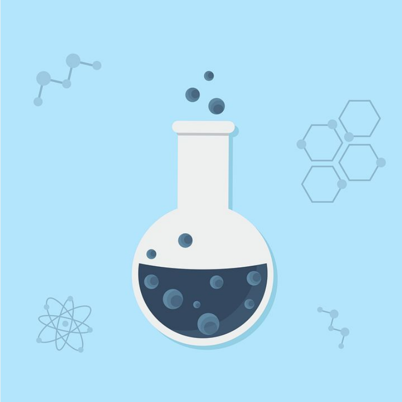
1. The NLP Reproducibility Checklist
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

OVERVIEW

1. **The NLP Reproducibility Checklist**
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

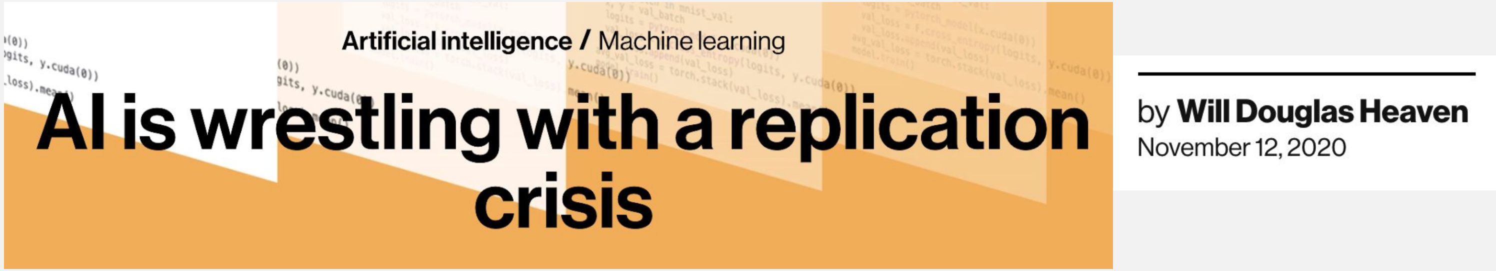
REPRODUCIBLE SCIENCE IS HARD

ML and NLP are driven by experiments.



The most important idea: reporting!

REPRODUCIBLE SCIENCE IS HARD



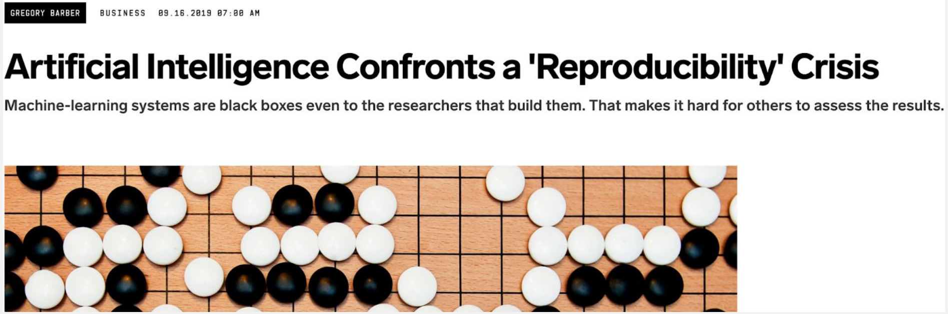
Artificial intelligence / Machine learning

AI is wrestling with a replication crisis

by **Will Douglas Heaven**
November 12, 2020

The Importance of Reproducibility in Machine Learning Applications

Home | The Importance of Reproducibility in Machine Learning Applications

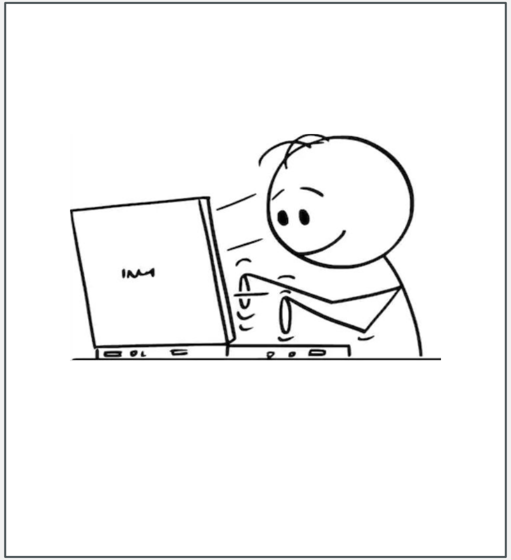


GREGORY BARBER   BUSINESS   09.16.2019 07:00 AM

Artificial Intelligence Confronts a 'Reproducibility' Crisis

Machine-learning systems are black boxes even to the researchers that build them. That makes it hard for others to assess the results.

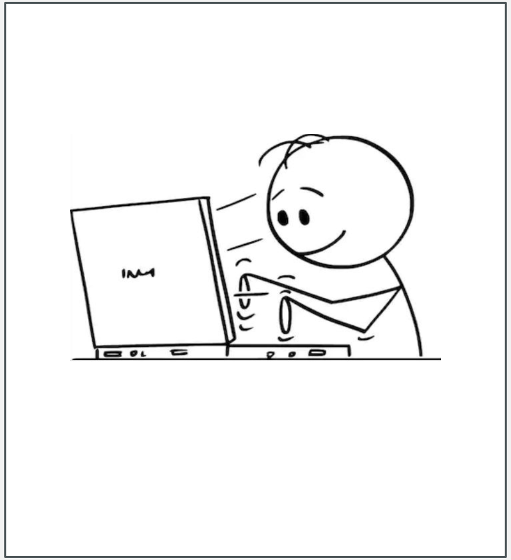
Get an idea



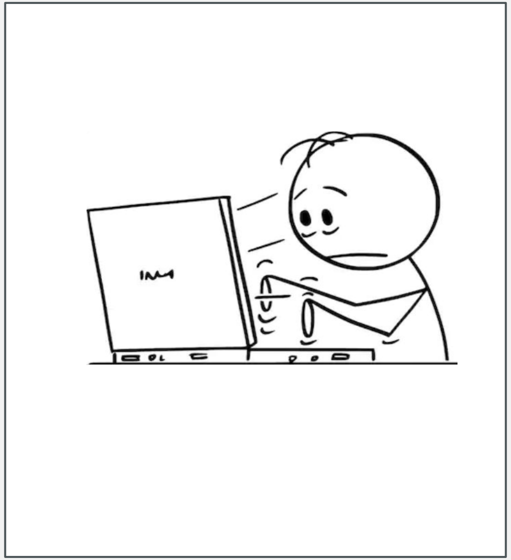
Get an idea

Spend time working

Get an idea



Spend time working

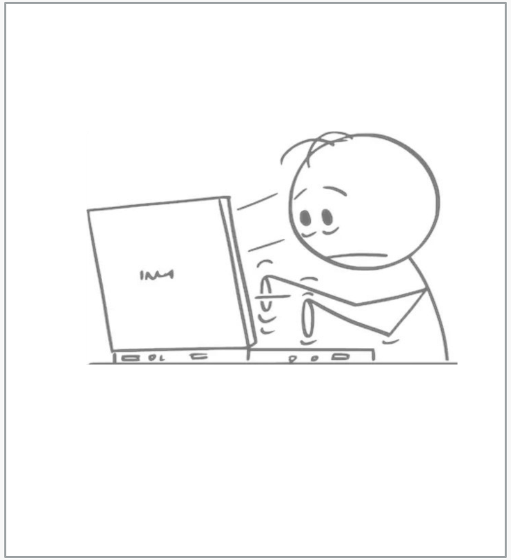


Negative result wasn't reported

Spurious correlation

Get an idea

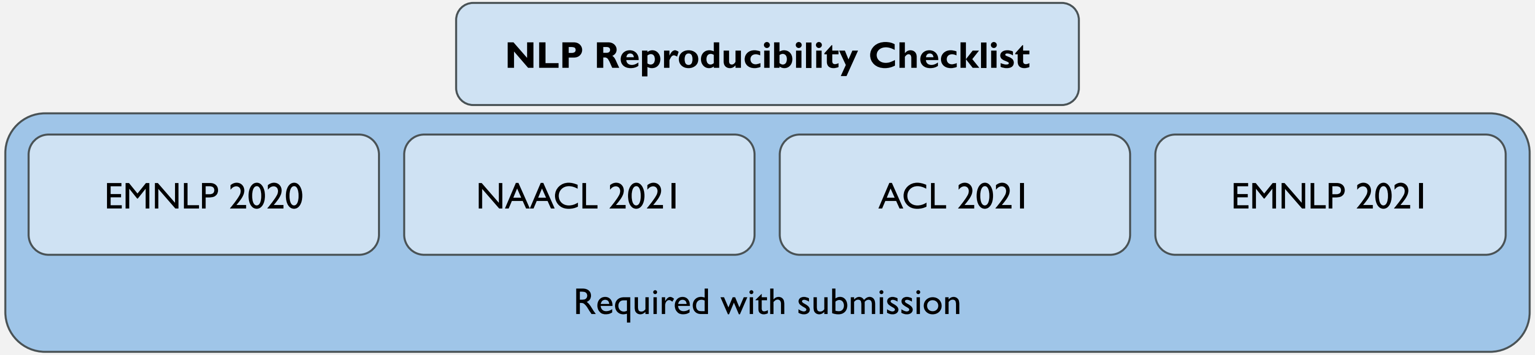
Spend time working



Negative result wasn't reported

Spurious correlation

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!



NLP Reproducibility Checklist

EMNLP 2020

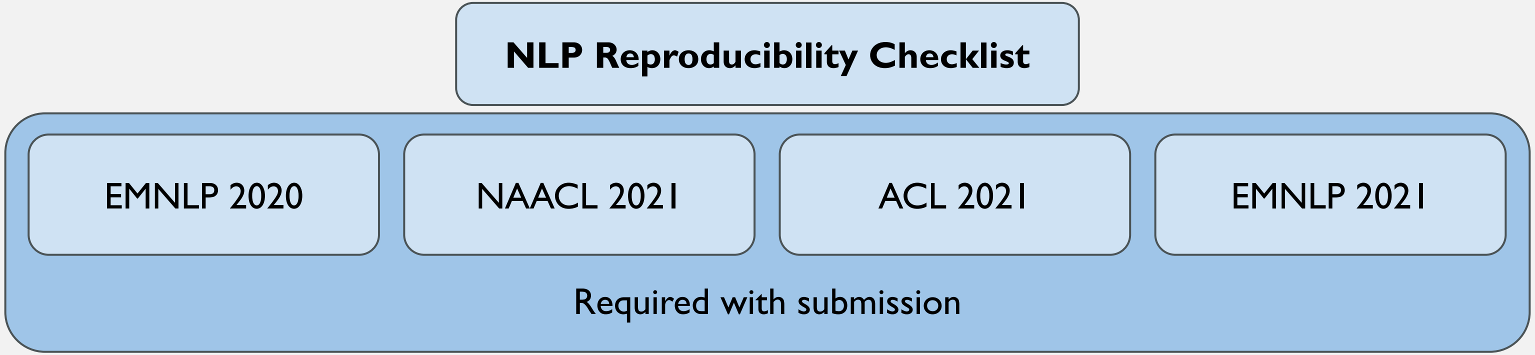
NAACL 2021

ACL 2021

EMNLP 2021

Required with submission

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!



NLP Reproducibility Checklist

EMNLP 2020

NAACL 2021

ACL 2021

EMNLP 2021

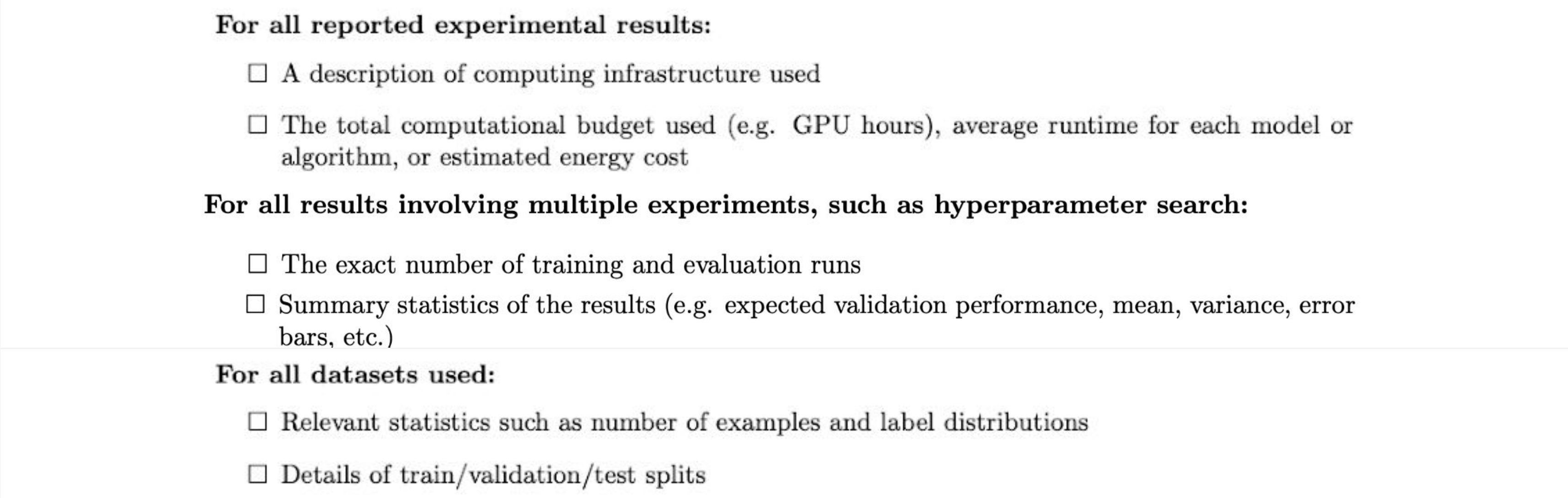
Required with submission

More than 10,000 submissions filled it out!

Goal: Remind authors of what they know they should report

HOW TO DO REPRODUCIBLE SCIENCE? REPORT ALL THE INFO YOU HAVE!

Example items:



For all reported experimental results:

- □ A description of computing infrastructure used
- □ The total computational budget used (e.g. GPU hours), average runtime for each model or algorithm, or estimated energy cost

For all results involving multiple experiments, such as hyperparameter search:

- □ The exact number of training and evaluation runs
- □ Summary statistics of the results (e.g. expected validation performance, mean, variance, error bars, etc.)

For all datasets used:

- □ Relevant statistics such as number of examples and label distributions
- □ Details of train/validation/test splits

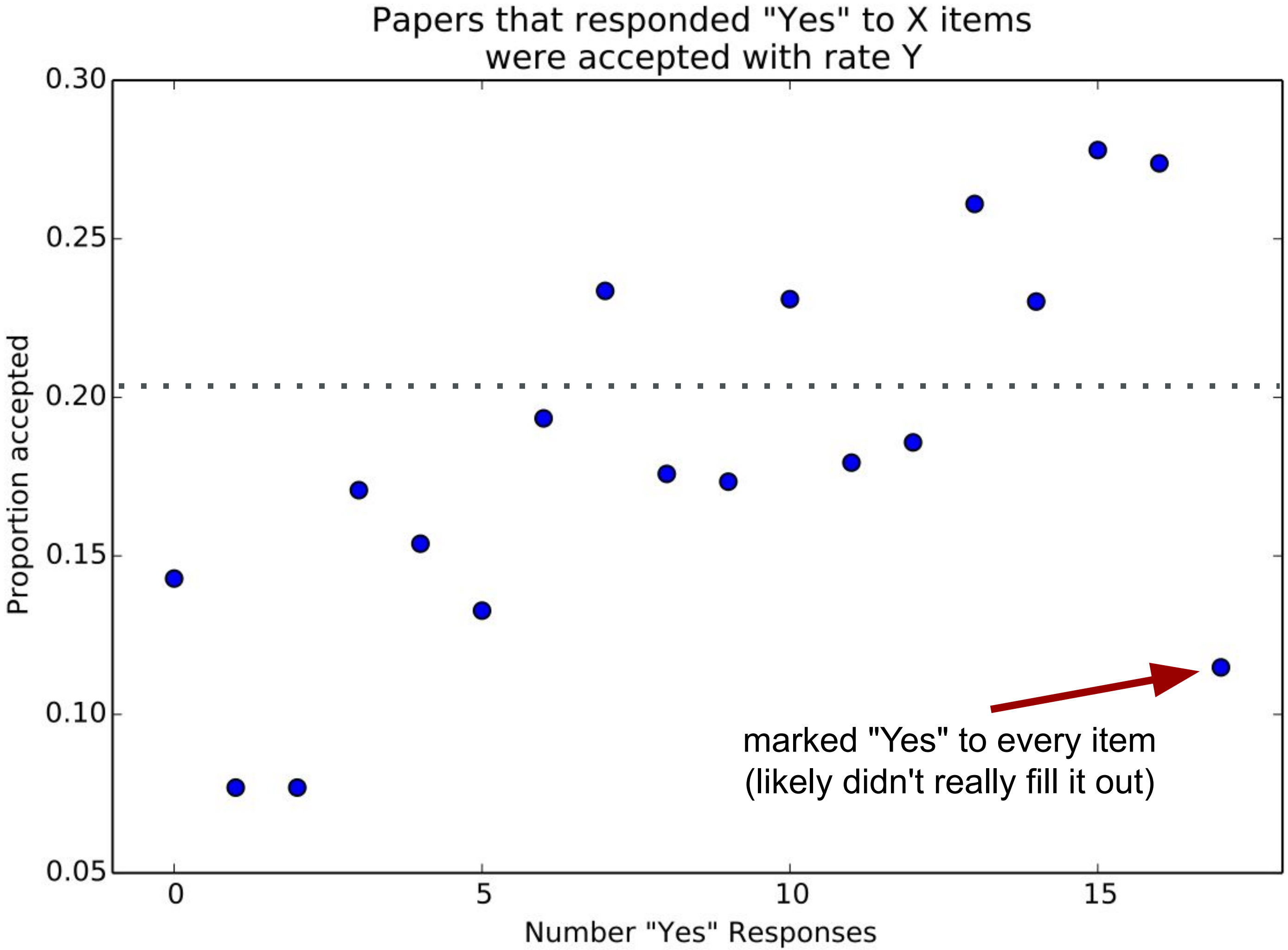
NLP REPRODUCIBILITY CHECKLISTS RESULTS - FIRST LOOK

EMNLP 2020

Total submissions: 3,677

Total accepted: 752 (20.4%)

First conference!
Likely different now



Papers that responded "Yes" to X items were accepted with rate Y

marked "Yes" to every item (likely didn't really fill it out)

READERS LIKE RELEVANT INFO

Items correlated with acceptance:

- "Average runtime for each approach"
- "Description of computing infrastructure used"

Room for improvement

- "Included all preprocessing steps" -- 11% marked "No"
- "Included a link to download the data" -- only 64% marked "Yes"

Data matters!

OVERVIEW

1. **The NLP Reproducibility Checklist**
2. The Responsible NLP Checklist
3. NLP Research on Reproducibility

OVERVIEW

1. The NLP Reproducibility Checklist
2. **The Responsible NLP Checklist**
3. NLP Research on Reproducibility

RESPONSIBLE NLP CHECKLIST

Required with submission to ARR

Combines Reproducibility + Ethics

Collaboration between NAACL PCs, ARR Editors, Anna Rogers, Margot Mieskes

Framed in terms of transparency: “Did you report [information]?”

Goal: Remind authors of what they know they should report

RESPONSIBLE NLP CHECKLIST

Required with submission to ARR

Combines Reproducibility + Ethics

Collaboration between NAACL PCs, ARR Editors, Anna Rogers, Margot Mieskes

Framed in terms of transparency: "Did you report [information]?"

Goal: Remind authors of what they know they should report

Marking "No" or "N/A" is not grounds for rejection!

RESPONSIBLE NLP CHECKLIST 1

- For every submission:
    - Describe limitations?
    - Describe risks?
    - Abstract and intro summarize main claims?

RESPONSIBLE NLP CHECKLIST 2

- Did you use or create scientific artifacts?
  - Cite creators?
  - Discuss the license or terms?
  - State intended use? Use consistently with creators intended use?
  - Personal information in new data?
  - Documentation of data?
  - Details of train / test / dev?

RESPONSIBLE NLP CHECKLIST 3

- Did you run computational experiments?
  - Number of parameters, total budget (e.g., GPU hours), computing infrastructure?
  - Hyperparameter search?
  - Error bars around results?
  - Details about software packages

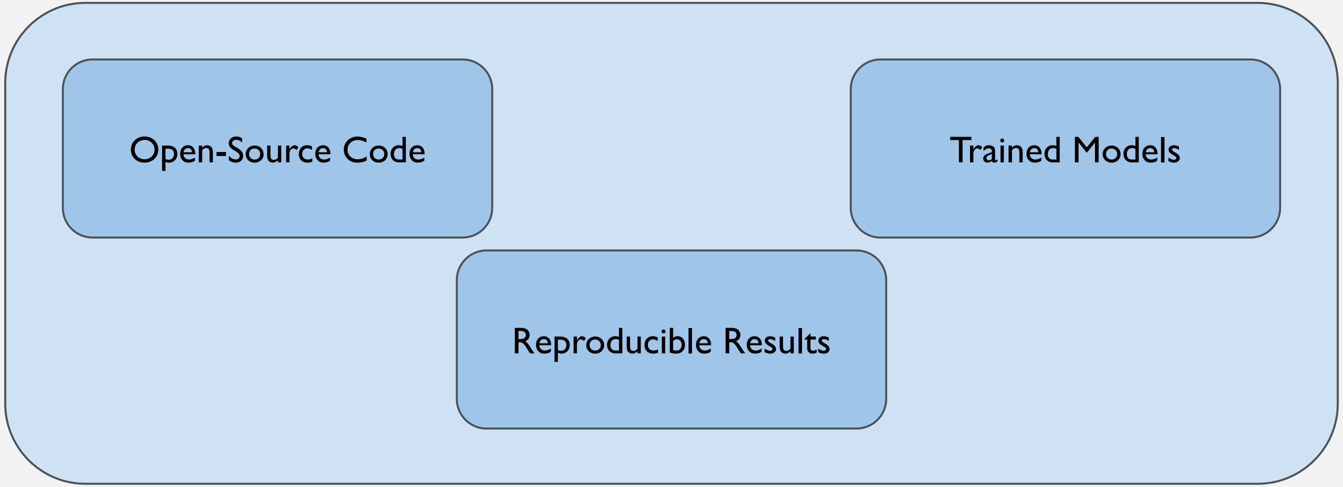
RESPONSIBLE NLP CHECKLIST 4

- Did you use human annotators (e.g., crowdworkers) or research with human participants?
  - Report the full text of instructions?
  - Report information about how you recruited, is payment adequate?
  - Did you get consent for intended use?
  - Approved by IRB?
  - Report the demographic info of annotators?

NLP CHECKLIST CONCLUSIONS

1. Report all the info you can!
2. The checklists are forward looking, cover best practices
3. You can mark "No" or "N/A" (with a good reason)

REPRODUCIBILITY CHAIR AT NAACL 2022

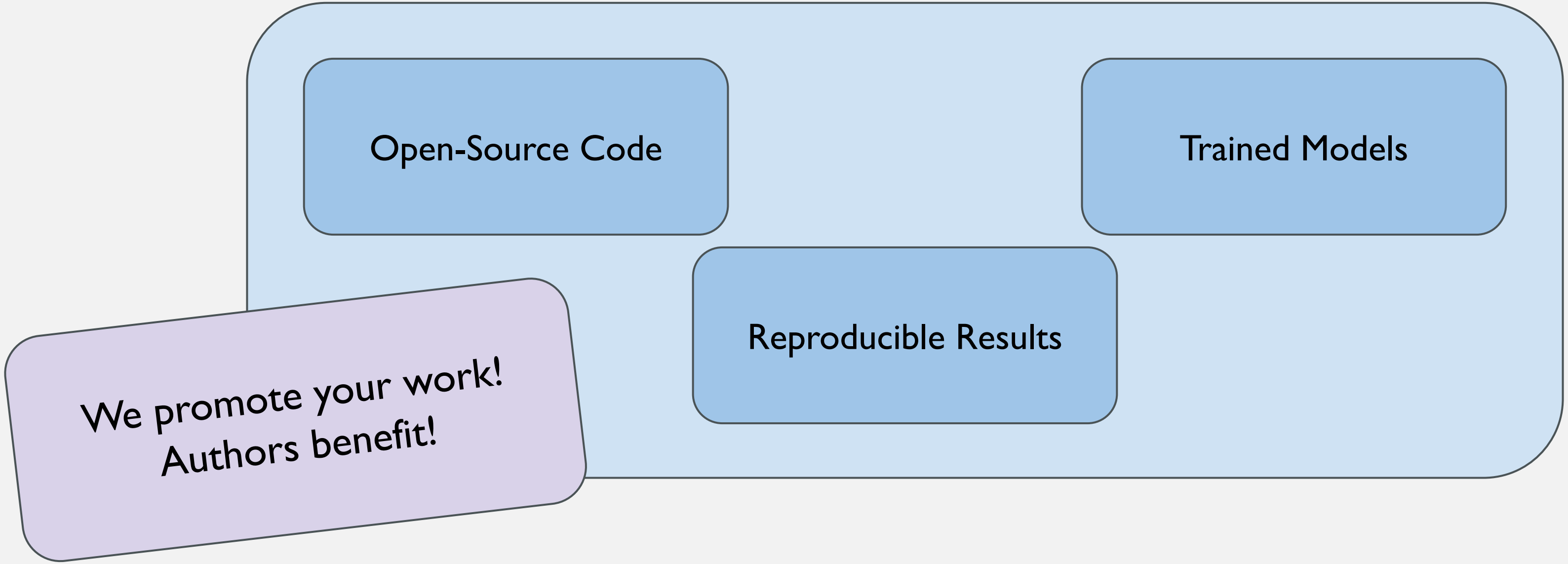


Open-Source Code

Trained Models

Reproducible Results

REPRODUCIBILITY CHAIR AT NAACL 2022



Open-Source Code

Trained Models

Reproducible Results

We promote your work!
Authors benefit!

OVERVIEW

1. The NLP Reproducibility Checklist
2. **The Responsible NLP Checklist**
3. NLP Research on Reproducibility

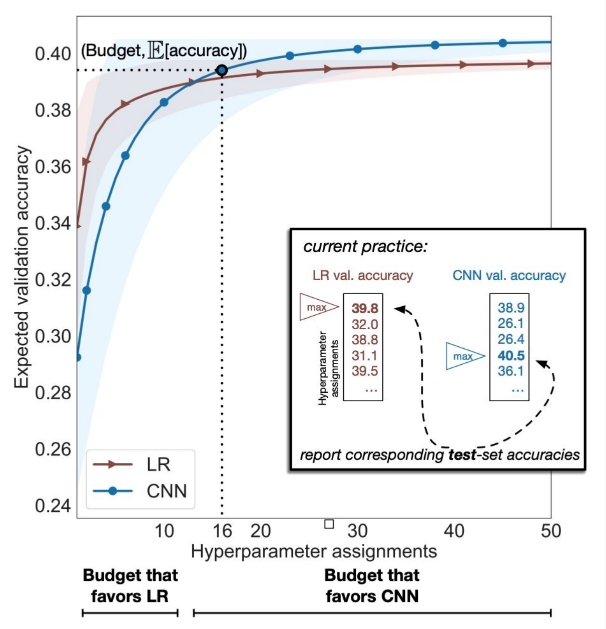
OVERVIEW

1. The NLP Reproducibility Checklist
2. The Responsible NLP Checklist
3. **NLP Research on Reproducibility**

REPORTING OF RESULTS (DODGE ET AL., 2019)

- Different budgets for hparam search lead to different conclusions about which model performs best

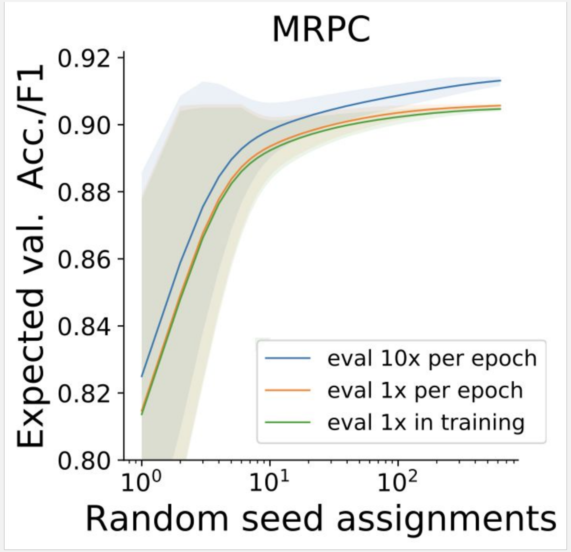
- Solution: report expected valid. perf.



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

RANDOM SEEDS (DODGE ET AL., 2020)

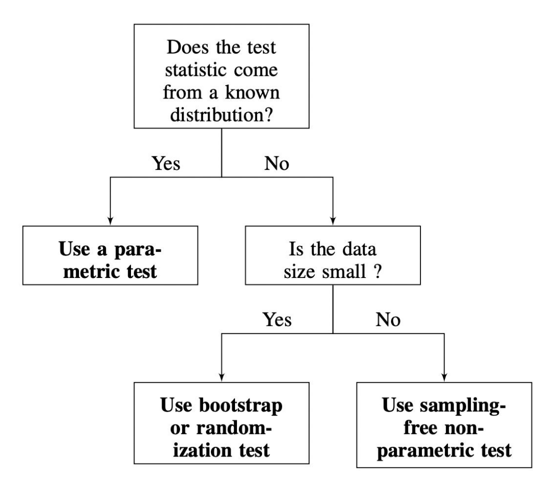
- Setup: fine-tuning BERT on GLUE tasks (MRPC, SST, CoLA, RTE)
- Conclusion: Surprisingly large variance from random seed!
- Suggestion: Start many runs, stop some early, report error bars



Dodge et al., Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. arXiv [cs.CL]. 2020.

STATISTICAL TESTING (DROR ET AL. 2018)

1. The authors provide a survey of which metrics are used of evaluation, how are metrics reported, and which statistical tests are used in NLP
2. The authors additionally provide a review of statistical tests that are relevant to NLP researchers and a flow chart to help them select a statistical test
3. The authors note that controlling for multiple hypothesis tests (Bonferroni correction) is also an important consideration when conducting statistical tests (Dror et al., 2017)


Does the test statistic come from a known distribution?
Yes
No
Use a parametric test
Is the data size small ?
Yes
No
Use bootstrap or randomization test
Use sampling-free non-parametric test

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

STATISTICAL POWER (CARD ET AL., 2020)

1. Statistical power measures how likely we will correctly reject the null hypothesis of a statistical test.
2. Card et. al analyze the statistical power of experiments in NLP
3. Many experiments lack statistical tests and sufficient statistical power.
4. Power analyses should be included as part of experimental planning.
    a. Experiments that cannot be conducted with sufficient statistical power may not lead to clear conclusions and should be carefully considered.
5. Code and model release, significance testing, and appropriate sample size can improve the quality of statistical analysis in the field

Card et al. With Little Power Comes Great Responsibility. EMNLP. 2020.

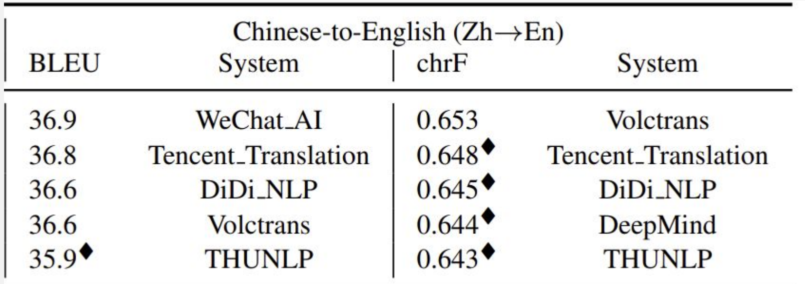
STANDARD SPLITS (GORMAN & BEDRICK, 2019)

1. Gorman & Bedrick compare utilizing the “standard split” of a provided dataset, versus randomly selecting the train, validation, and testing split for POS tagging task.
2. They utilize statistical testing as recommended in Dror et al. to correct for multiple hypotheses.
3. Many methods for POS are overfitted to the standard split and do not perform as well on a randomly generated split.
4. The authors recommend Bonferroni corrected random split hypothesis testing to confirm that results on the standard split are robust to random split

Gorman and Bedrick. We Need to Talk about Standard Splits. ACL. 2019

MACHINE TRANSLATION (MARIE ET AL., 2020)

- A large-scale meta-evaluation of Machine Translation (MT), manually annotating 769 research papers published from 2010-2020.
- It found several issues:
  - the exclusive use of BLEU, a metric with significant limitations
  - the absence of statistical significance testing
  - the comparison of incomparable results from previous work
  - comparing MT systems that do not exploit the same data
- Depending on the metric being used, different systems can be considered as superior.



Chinese-to-English (Zh→En)			
BLEU	System	chrF	System
36.9	WeChat_AI	0.653	Volctrans
36.8	Tencent_Translation	0.648♦	Tencent_Translation
36.6	DiDi_NLP	0.645♦	DiDi_NLP
36.6	Volctrans	0.644♦	DeepMind
35.9♦	THUNLP	0.643♦	THUNLP

Marie et al., . Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. ACL. 2021.

MACHINE TRANSLATION (MARIE ET AL., 2020)

Data differences also impact scores:

- Tokenizer used
- Dataset preprocessing (e.g. max length or language ID used for filtering)

The authors propose guidelines for automatic MT evaluation, including:

1. Metrics other than BLEU
2. Statistical significance testing
3. Reproducing previous scores instead of copying them
4. Ensuring that the data, splits, and preprocessing used are the same

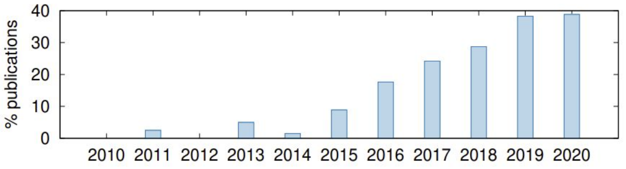


Figure 4: Percentage of papers that compared MT systems using data that are not identical.

Marie et al., . Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. ACL. 2021.

TRANSFORMERS (NARANG ET AL., 2021)

- An extensive evaluation of different Transformer modifications in a shared experimental setting in NLP:
  - Activations, normalization, depth, embeddings, parameter sharing, softmax, applied to different Transformer architectures
- They find that many Transformer modifications **do not** result in improved performance, and suggest that changes to Transformers suffer from lack of generalization across different implementations and tasks.
- The authors also found that **hyperparameter tuning** was a major challenge for Transformers given the space of possible combinations

Narang et al. Do Transformer Modifications Transfer Across Implementations and Applications? EMNLP. 2021.

TRANSFORMERS (NARANG ET AL., 2021)

Some proposals made to ensure the robustness of improvements include:

- Trying changes out in multiple codebases
- Applying them to a wide variety of downstream applications, including domains outside of NLP
- Keeping hyperparameters fixed as much as possible, and/or measuring the robustness of the modifications to changes in hyperparameters
- Reporting of results should include mean and standard deviation across multiple trials

Narang et al. Do Transformer Modifications Transfer Across Implementations and Applications? EMNLP. 2021.

REPRODUCIBILITY IN LARGE LANGUAGE MODELS

1. Models such as Transformer XL, Megatron, GPT-Neo, OPT, T0 share code on GitHub
2. Big Science and OPT share model logs
3. Open datasets such as OpenWebText and the Pile aid in pretraining
4. HuggingFace provides a model zoo of pre-trained weights (many shared by the original authors)
5. Checkpoints and replicates such as MultiBert enable researchers to study training dynamics and variability
6. Tools such as evaluationharness, promptsource, codecarbon provide useful evaluation

Ensuring that our research is reproducible remains an important goal within NLP research

But it is not the only consideration

LIMITATIONS IN REPRODUCIBLE NLP

1. Environmental Impact
2. Depreciation of hardware/software
3. Ethical Challenges

ENVIRONMENTAL IMPACT

1. Reproducing papers from scratch creates additional environmental cost
2. Sharing models and hyperparameters makes it possible to avoid these costs
3. Clearly communicating energy requirements and carbon emissions also makes it possible to take these into account when choosing between different models
4. Tools such as codecarbon, Azure has an upcoming tool[1]. Allow for calculations of carbon emissions

https://techcommunity.microsoft.com/t5/green-tech-blog/charting-the-path-towards-sustainable-ai-with-azure-machine/ba-p/2866923

DEPRECATION

1. Operating under assumption that we're using the same hardware and software as the original paper, which becomes less likely as time goes on [Mesnard & Barba, 2017]
2. Researchers are not incentivized to maintain their code
3. Dataset deprecation:
   - Versions: e.g. Common Crawl, Wikipedia get updated regularly
   - Datasets removed by creators: TinyImages, Duke MTMC, etc. – but continue being used
   - No centralized identification schema for datasets (e.g. DOI)
   - Current endeavors by conferences like NeurIPS are aiming to create a centralized repository for deprecated datasets

Mesnard and Barba. Reproducible and Replicable Computational Fluid Dynamics: It's Harder Than You Think. Computing in Science Engineering. 2017.

ETHICAL CHALLENGES

1. Reproduction of NLP papers does not happen in a vacuum - considerations when conducting reproduction studies should also take into account the ethical considerations particular to a given methodology
2. The ACL Code of Ethics and ACL Rolling Review Responsible NLP Research checklist provide a useful starting point to help researchers conduct and share their work responsibly
3. Misunderstanding a paper can lead a researcher to make incorrect assumptions when reproducing the paper

MOTIVATION

How can we mitigate the challenges of bigger, more complex models without reducing the benefits?

In this tutorial, we focus on the challenge of ensuring research results are reproducible

TUTORIAL OVERVIEW

1. Introduction to Reproducibility
2. Reproducibility in NLP
3. Mechanisms for Reproducibility
4. Reproducibility as a Teaching Tool