THE REPROGEN SHARED TASK ON REPRODUCIBILITY OF EVALUATIONS IN NLG




Anya Belz, ADAPT/DCU, Ireland
Anastasia Shimorina, Orange, France
Maja Popović, ADAPT/DCU, Ireland
Ehud Reiter, University of Aberdeen, UK

Overview

1. Background
2. ReproGen Shared Task Overview
3. Participating Teams
4. Results
5. Discussion and Conclusion

Background

- Experimental results should be reproducible!
  - ACM: “An experimental result is not fully established unless it can be independently reproduced.”
- Some work in NLP on reproducing metric-based exper
- NLG use a lot of human evaluation, but little is known about reproducibility of human eval
  - Cooper & Shardlow (2020): reproduced system rankings in text simplification, but reported ~15% lower mean scores
  - Belz & Kow (2010): Pearson’s 0.84-0.99 when same evaluators are used (data-to-text, REG)

ReproGen

- Shared task on reproducing evaluations of NLG sys
  - Get real data on reproducibility!
- Two tracks
  - Track A - Main Reproducibility Track: Reproduction of selected papers. I.e. participants repeat an experiment and report results
  - Track B - RYO Track (Reproduce Your Own): Reproduction of own results. As for Track A, but for participants' own results
  - For human eval, Human Evaluation Data Sheet (HEDS) required
  - Light touch review and feedback to participants

ReproGen 2021 and 2022

- ReproGen 2021 (INLG 2021 Generation Challenges)
  - 4 submissions, 2 in each track
  - Just human evaluations

- ReproGen 2022 (INLG 2022 Generation Challenges)
  - 5 submissions, 3 in track A, 2 in track B
  - Both automatic and human evaluations

ReproGen 2022: Choosing papers in main track

- CFP for papers in Track A; with inclusion criteria:
  - Paper must include information necessary to repeat an evaluation, OR authors can provide it
  - Authors commit to being available during reproduction in case of questions
  - Evaluation has to be repeatable in principle, and be low-cost
  - Has to be on NLG task with text as output
- Selected 5 papers: 4 from ReproGen 2021 one new one

Shared Task Overview

- Selected 5 papers:
  - **van der Lee et al. (2017)**: PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences: 1 evaluation study; Dutch; 20 evaluators; 1 quality criterion; reproduction target: primary scores
  - **Dušek et al. (2018)**: *Findings of the E2E NLG Challenge*: 1 evaluation study; English; MTurk; 2 quality criteria; reproduction target: primary scores
  - **Qader et al. (2018)**: *Generation of Company descriptions using concept-to-text and text-to-text deep models*: dataset collection and systems evaluation: 1 evaluation study; English; 19 evaluators; 4 quality criteria; reproduction target: primary scores
  - **Shaikh & Santhanam (2019)**: *Towards Best Experiment Design for Evaluating Dialogue System Output*: 3 evaluation studies differing in experimental design; English; 40 evaluators; 2 quality criteria; reproduction target: correlation scores between 3 studies
  - ***(New paper for 2022) Nisioi et al. (2017)***: *Exploring Neural Text Simplification Models* : one automatic evaluation study; reproduction target: two automatic scores; one human evaluation study; 70 sentences; 9 system outputs; 4 quality criteria; reproduction target: primary scores

Participating Teams

Track	Team	Original paper	Reproduction paper	Metrics
A	Tilburg University	Santhanam and Shaikh (2019)	Braggaar et al. (2022)	automatic, human
A	ADAPT @ DCU	Nisioi et al. (2017)	Popović et al. (2022)	human
A	Univ of Illinois, Chicago	Nisioi et al. (2017)	Arvan et al. (2022)	automatic
B	Univ of Aberdeen	Thomson and Reiter (2021) [annotate errors]	Thomson and Reiter (2022)	human
B	ADAPT+CU+Minas Gerais	Dušek and Kasner (2020) [NL inference]	Huidrom et al. (2022)	human

8

Results

I. Compare original vs. reproduction score sets:

1. Pearson's r
2. Spearman's rho
3. Mean percentage increase/decrease
4. Mean absolute percentage change
5. Mean coefficient of variation (CV), a standard measure of precision in metrology
    - CV = standard deviation over the mean
    - CV* corrects for small sample size by using sample standard deviation calculated from the *unbiassed* sample variance

Results: Track A

measurand(s)	Pearson's rho	Spearman's r	Mean change +/-	Mean change abs	Mean CV*
Orig study = Nisioi et al. (2017); reproduction (metric eval) = Arvan et al. (2022); same outputs					
All Scores (2 sys×2 metrics)	1	1	0	0	0
Orig study = Nisioi et al. (2017); reproduction (metric eval) = Arvan et al. (2022); regen outputs from code					
All Scores (2 sys×2 metrics)	1	0.8	-1.02	3.30	3.34
Original study = Nisioi et al. (2017); reproduction (metric eval) = Arvan et al. (2022); corrected code					
All Scores (2 sys×2 metrics)	1	0.8	0.63	3.19	3.16
Orig study = Nisioi et al. (2017); reproduction (human eval) = Popović et al. (2022); different evaluators					
All Scores (9 sys1 qual crit)	0.766	0.787	40.16	85.82	8.98
Orig study = Santhanam and Shaikh (2019); reproduction (human/metric)= Braggaar et al. (2022)					
All Scores (2 corr coeff×2 qual crit×4 scales)	0.01	0.16	9.26	12.71	11.70

10

Results: Track B

measurand(s)	Pearson's rho	Spearman's r	Mean change +/-	Mean change abs	Mean CV*
Orig study = Dušek and Kasner (2020); repro (human eval) = Huidrom et al. (2022); same evaluators					
All Scores (8/9 label counts × 1 system × 2 datasets)	0.81	0.87	20.12	43.00	34.34
Orig study = Dušek and Kasner (2020); repro (human eval) = Huidrom et al. (2022); different evaluators					
All Scores (8/9 label counts × 1 system × 2 datasets)	0.84	0.66	18.76	48.79	39.16
Orig study = Thomson and Reiter (2021); repro (human eval) = Thomson and Reiter (2022); diff data set					
All Scores (6 label counts × 3 sys)	0.89	0.88	33.6	60.10	52.75
Orig study = Thomson and Reiter (2021); repro (human eval) = Thomson and Reiter (2022); diff data set					
All Scores (6 label counts × 3 sys)	0.896	0.84	30.12	77.06	68.00

Discussion

- Degree of reproducibility better for automatic metric
- Repro of human eval had decent corr but high CV
  - Agreed on which system was “best”
  - Did not agree on specific score of system
  - Human “noise” affect absolute scores more than rankings?
- Hard to make strong claims
  - Only 5 reproductions, small experiments
  - Simple evaluations have better degree of reproducibility??

Conclusion

- NLG evaluations need to be reproducible!
  - a rerun of experiment should produce similar results
- Need experimental evidence which shows which studies are reproducible and which are not
  - Ideally create guidelines based on results
- ReproGen is contribution towards this goal
- Also ReproHum project (https://reprohum.github.io/)