M.INF.2801 RESEARCH LAB ROTATION

Meh-Tricks: Towards Reproducible Results in NLP

Christopher L. Lübbers
Supervisor: Dr. Terry Lima Ruas
2024-03-15

Agenda

1. Introduction
2. Methodology
3. Results
4. Conclusion

The Importance of Accurate Evaluation in NLP

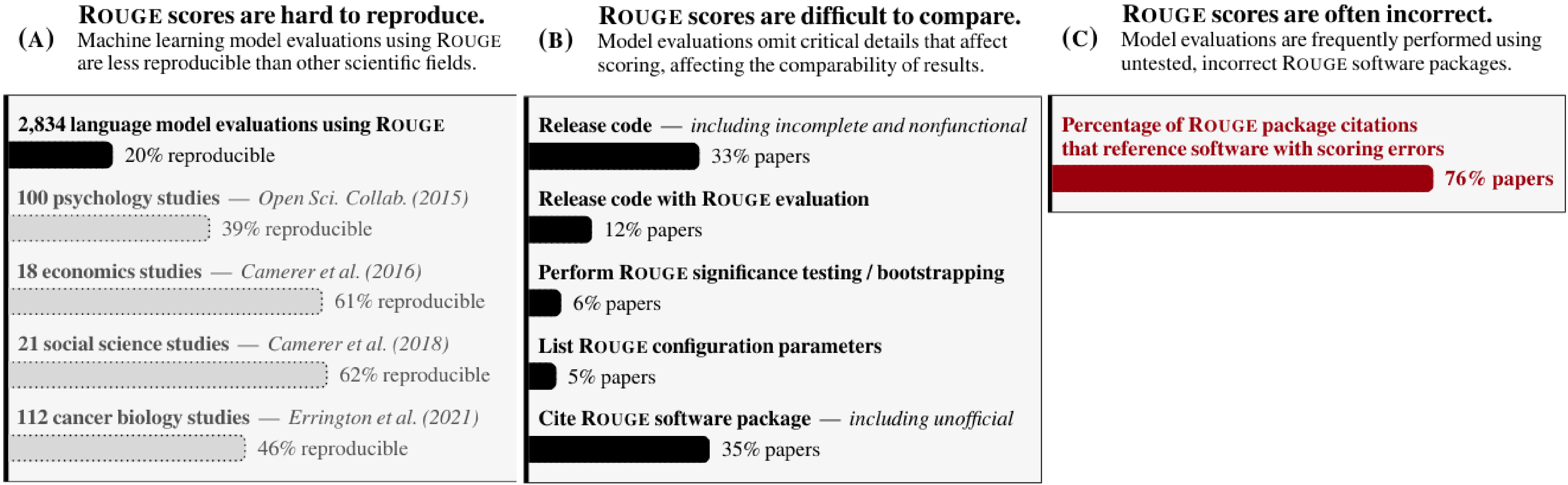
- Progress in NLP is often measured in improvement of performance metrics
- Scores and leaderboards influence both academic research and practical applications
- Key metrics in NLP: **ROUGE**, METEOR, BLEU

Many ROUGE configuration differences are bigger than leaderboard model differences.

Common ROUGE Configurations	Change in ROUGE Scores (Compared to Baseline Config.) ± R1	± R2	± RL
Preprocessing			
Apply Stemming	+1.68	+0.54	+1.31
Remove Stopwords	–2.21	–0.58	–0.99
Tokenization			
No Sent. Splits	[Sent. splits have no effect on ROUGE-N]		–11.17
Period Sent. Splits			–3.44
NLTK Sent. Splits			–0.16
NLTK Tokenize	<0.01	<0.01	<0.01
Truncation (Recall)			
Truncate to 75 Bytes	–27.92	–12.93	–33.44
Truncate to 100 Words	–0.07	–0.05	–0.07
Misreported Scores			
Report $F_{1.2}$ Score	+1.33	+0.61	+1.21
Report Recall Score	+10.88	+5.00	+9.92
Helpful Comparison The average ROUGE score difference between the current top five CNN / Daily Mail models.	**±0.50**	**±0.18**	**±0.53**

[Rogue Scores](Grusky, ACL 2023)

Challenges in Current NLP Evaluation Practices



ACL Anthology + DBLP = 110,689 papers by January 2023

Rogue Scores (Grusky, ACL 2023)

Objective

- RQ: To what extent do variations in methodologies and libraries of METEOR lead to inconsistencies in reported results?
- Tasks
  - Literature Review
  - Analysis of Metric
  - Baseline Evaluation

Methodology

Gipp Lab
Scientific Infomation Analytics

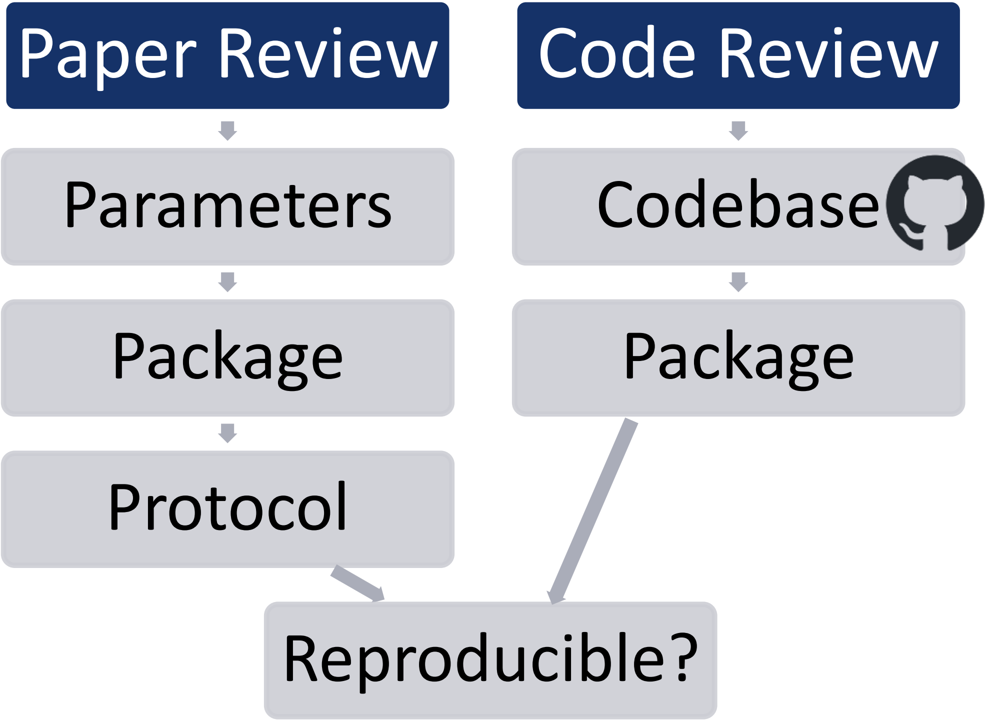
Systematic Literature Review

Jupyter Notebook

ACL Anthology Dataset

(September 2022)

- METEOR identification



Logos: https://github.com/jupyter/jupyter.github.io/blob/main/assets/logos/logomark-orangebody-greyplanets.svg
https://github.com/acl-org/acl-anthology/blob/master/hugo/static/images/acl-logo.svg
https://github.com/logos

Details of Review

- Iteratively choosing packages
- Manual Code review
- Limitations
  - No manual paper review
  - 1 venue, peer-reviewed papers, current versions, no external material
  - Automated annotation, preliminary search, only English, no author clarification, non-evaluation metrics, assumed wrapper correctness
  - Codebase linking, multiple packages
  - only GitHub repositories

Software Validation Testing

- Docker image
- Package implementation
- Task: CNN /Daily Mail dataset
  - Model: evaluate Lead-3
  - Mean of 13k METEOR scores

Logos: https://www.docker.com/company/newsroom/media-resources/
https://huggingface.co/brand

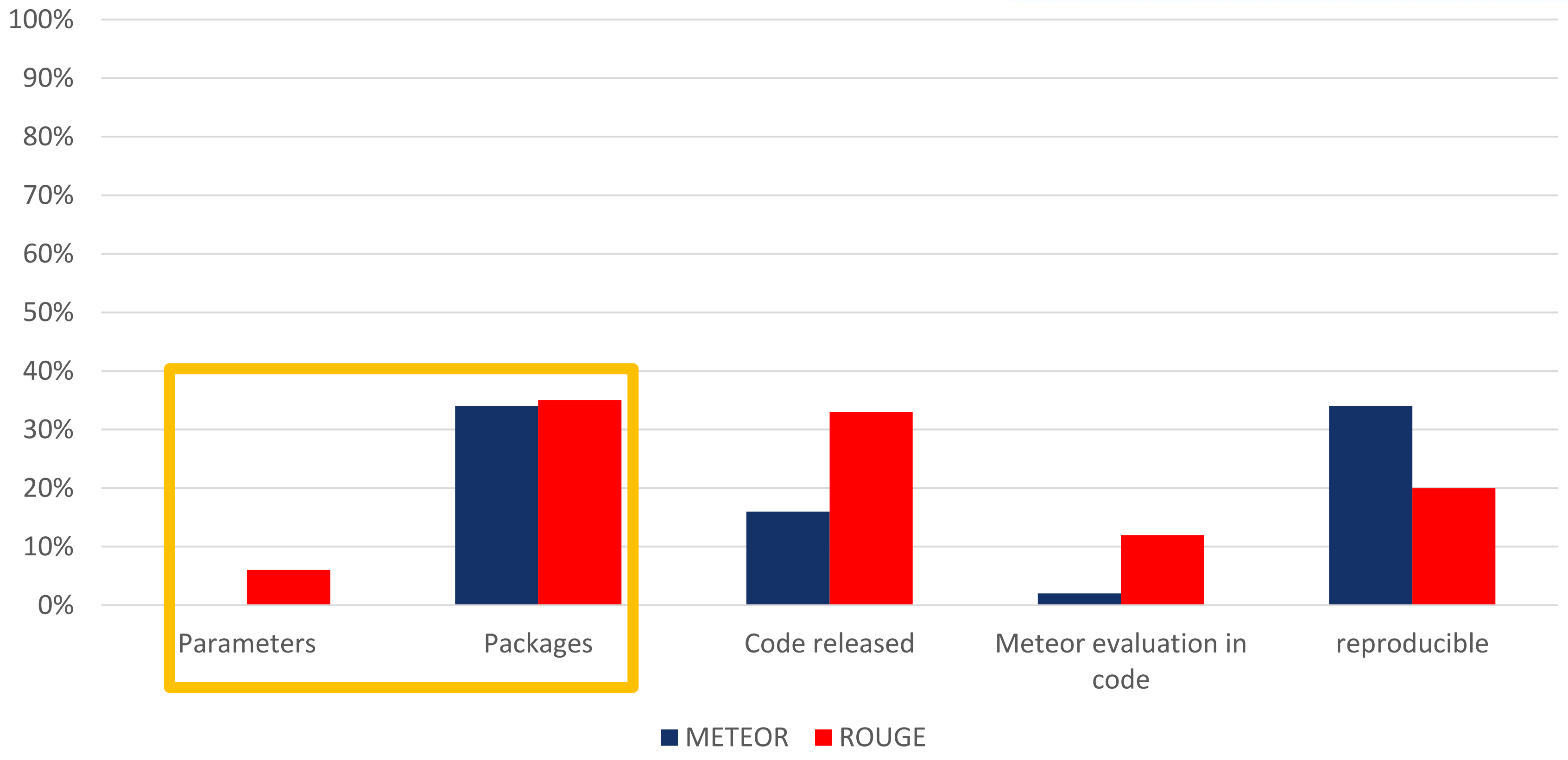
Package implementation

- Dependencies
- different input formats
- METEOR only part of a bigger suite
- Limitations
    - Only packages compatible with Python3 tested
    - Single task, only English evaluation, no multiple references, package versions

Results

Gipp Lab
Scientific Infomation Analytics

Reproducibility of 1613 papers



Correctness

Github	METEOR score	% of citations
salaniz/pycocoevalcap	0.218336	60 %
WebNLG/GenerationEval	0.217303	
Maluuba/nlg-eval	0.221483	
Yale-LILY/SummEval	0.221483	
nltk/nltk	0.382306	40 %
huggingface/evaluate	0.382306	
facebookresearch/vizseq	0.332000	

+75 %

Influence of Parameters on Scores

	METEOR	Alpha	Beta	Gamma	Delta	Weights	Score
Official	1.0	0.8	2.5	0.4	False	False	-
NLTK default	1.0	0.9	3.0	0.5	False	False	0.382306
Official	1.5	0.85	0.2	0.6	0.75	True	0.218336
NLTK	1.0	0.85	0.2	0.6	False	False	0.218184

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

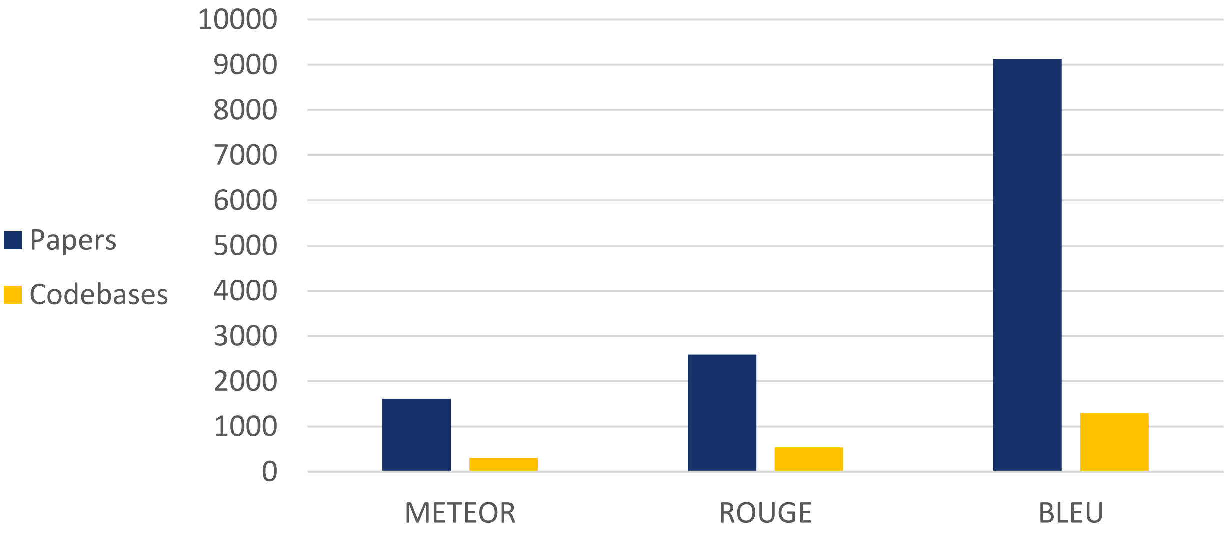
Summary of Results

- To what extent do variations in methodologies and libraries of METEOR lead to inconsistencies in reported results?
  - Low reproducibility of papers (34 %)
  - Primary issue: wrapper vs reimplementation (difference of 75 %)
  - 40 % of reported scores are wrong

Conclusion

Outlook / Future Work

- Manual paper reviews
- Improve NLTK
- Review of BLEU scores



Conclusion

- Reproducibility
  - Aim for reproducibility
  - Reference used software
- Correctness
  - Know your metrics!
  - If you are building on another paper, make sure the metrics are comparable.
- METEOR
  - Strength: tuned default parameters
  - Weakness: popular packages use old parameters
  - Use: salaniz/pycocoevalcap
  - NLTK: ???

References

[1]
Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.

[2]
Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014. Association for Computational Linguistics, Baltimore, Maryland, USA, 376–380. https://doi.org/10.3115/v1/W14-3348

[3]
Michael Denkowski and Alon Lavie. Extending the METEOR Machine Translation Evaluation Metric to the Phrase Level.

[4]
Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems.

[5]
Max Grusky. 2023. Rogue Scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023. Association for Computational Linguistics, Toronto, Canada, 1914–1934. https://doi.org/10.18653/v1/2023.acl-long.107

[6]
Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, 2007. Association for Computational Linguistics, Prague, Czech Republic, 228–231. https://doi.org/10.3115/1626355.1626389

METEOR calculation

$$P = \frac{\sum_i w_i \cdot (\delta \cdot m_i(h_c) + (1-\delta) \cdot m_i(h_f))}{\delta \cdot |h_c| + (1-\delta) \cdot |h_f|}$$

$$R = \frac{\sum_i w_i \cdot (\delta \cdot m_i(r_c) + (1-\delta) \cdot m_i(r_f))}{\delta \cdot |r_c| + (1-\delta) \cdot |r_f|}$$

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

$$Score = (1 - Pen) \cdot F_{mean}$$

NLTK meteor_score.py

```
        alpha: float = 0.9,
        beta: float = 3.0,
        gamma: float = 0.5,
    ) -> float:
        """
        Calculates METEOR score for single hypothesis and reference as per
        "Meteor: An Automatic Metric for MT Evaluation with HighLevels of
        Correlation with Human Judgments" by Alon Lavie and Abhaya Agarwal,
        in Proceedings of ACL.
        https://www.cs.cmu.edu/~alavie/METEOR/pdf/Lavie-Agarwal-2007-METEOR.pdf
```

https://github.com/nltk/nltk/blob/develop/nltk/translate/meteor_score.py