Towards Inferential Reproducibility of Machine Learning Research

Michael Hagmann, Philipp Meier and Stefan Riezler

Computational Lingustics & IWR
Heidelberg University, Germany

{hagmann, riezler}@cl.uni-heidelberg.de

Aim

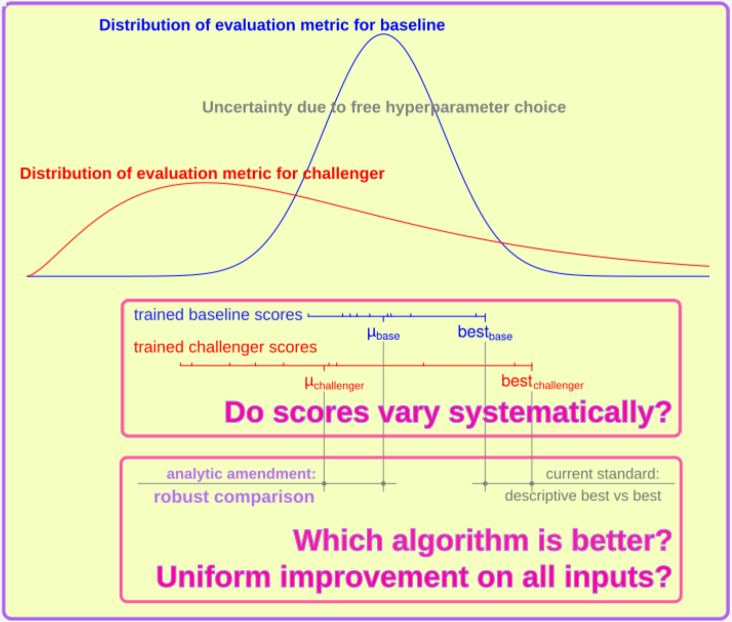
Compare (at least two) competing algorithms.

Training

- Algorithms have several hyper-parameters.
- Hyper-Parameters needed to be set before test-set optimization.
- Best model (hyper-parameter configuration) is found via dev-set performance ranking.

Evaluation

- Evaluation data: Best models are applied on test-set.
- These outputs are used to estimate expected out-of-sample risk.
- Descriptive comparison of the obtained estimates.


Distribution of evaluation metric for baseline
Uncertainty due to free hyperparameter choice
Distribution of evaluation metric for challenger
trained baseline scores
μbase
bestbase
trained challenger scores
μchallenger
bestchallenger
Do scores vary systematically?
analytic amendment:
robust comparison
current standard:
descriptive best vs best
Which algorithm is better?
Uniform improvement on all inputs?

Extended Evaluation Data

- All model instances for each algorithm are applied on test-set
- Record hyper-parameter values of each model
- Optionally: Add input characteristics of test examples

Linear Mixed Effect Models (LMEMs)

- Allow: Estimation of expected out-of-sample risks and differences
- Provide: Distribution for these estimates (via ML-Theory)
  - Enables statistical inference
- Allow: Analysis of non-iid (evaluation) data
- Allow: Complex conditional assessment of out-of-sample risk
- Allow: Assessment of variance components

R3F fine-tuning [Aghajanyan et al., 2021]

$$\mathcal{L}(\theta) + \lambda KL_{sym}(g \circ f_\theta(x) || g \circ f_\theta(x + z))$$
$$\text{s.t} \quad z \sim \mathcal{N}(0, \sigma^2 I) \text{ or } z \sim \mathcal{U}(-\sigma, \sigma)$$

Experiment

- Task: text summarization with BART
- Data: CNN/DailyMail [Hermann et al., 2015] and RedditTIFU [Kim et al., 2019]
- Evaluation metric: Rouge-1/2/L [Lin and Hovy, 2003]

Obstacles

- Data split for Reddit not provided; used split of [Zhong et al., 2020]
- RNG seeds not provided

CNN Data Set

	BART	BART-R3F	p-value	effect size
R-1	44.15 (44.16)	44.72 (44.38)	< 0.0001	-0.101
R-2	21.13 (21.28)	21.17 (21.53)	< 0.0001	-0.080
R-L	40.81 (40.90)	41.40 (41.17)	< 0.0001	-0.105

- **Significant but small** performance gain for BART-R3F.

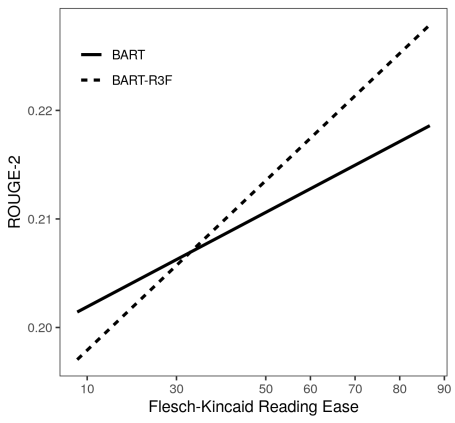
Measuring difficulty of summarization data

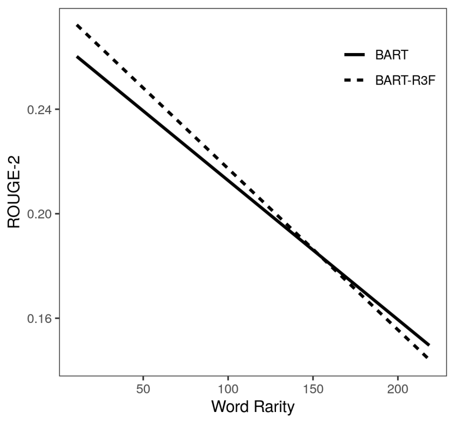
- **Word rarity** [Platanios et al., 2019]: Sum of negative log of empirical probabilities of words in segment.

  Higher value means harder data example.

- **Flesch-Kincaid readability** [Kincaid et al., 1975]: Index based on words/sentences and syllables/word; in principle unbounded, but interpretation scheme exists for ranges from 0 (difficult) to 100 (easy).

  Lower value means harder data example.





- Performance gains are not uniform across readability/word rarity.
- BART-R3F is only better on easy inputs.

Robust comparison (with meta-parameter variation)

- BART: 18 models (seeds)
- BART-R3F: 30 models (3 λ values, 2 noise distributions & 5 seeds)

CNN Data Set

	BART	BART-R3F	*p*-value	effect size
R-1	44.15	41.06	< 0.0001	0.384
R-2	21.30	19.00	< 0.0001	0.308
R-L	40.84	36.40	< 0.0001	0.500

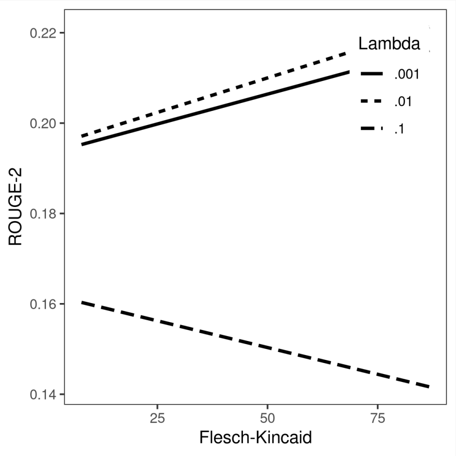
- **Traditional fine-tuning is better than R3F!**
- ➡ Detailed analysis of BART-R3F models!

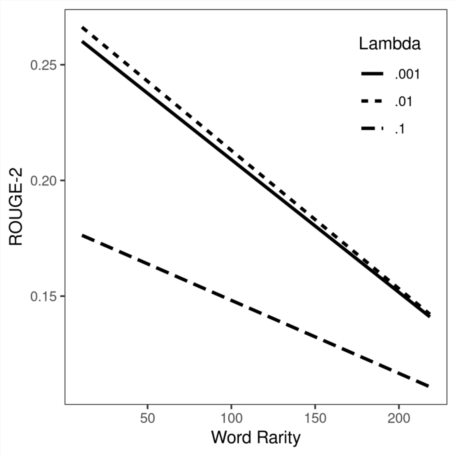
Variance Decomposition of Rouge-2 Scores

Source	Variance component	Percent
summary-id	**0.00992**	**62.70**
lambda	0.00131	8.31
random-seed	0.00008	0.48
noise-distribution	0.00003	0.20
residual	0.00449	28.3

- Only moderate reliability.
- Largest variance component for λ.

Interaction of Meta-Parameters with Data Properties





- Performance drop of BART-R3F for $\lambda = 0.1$.

Reproducibility Results on RedditTIFU

- Difficult data (mean readability score of -348.9).

- best vs best: BART-R3F only better for Rouge-2 at small effect size.
- robust comparison: No significant improvements.
- Reliability coefficients of $\approx 80\%$.
- λ variance component is negligible.

- Improvements of BART-R3F strongly depends on finding the **sweet spot of a single meta-parameter** (here: λ) – paper's goal was explicitly to reduce instability across meta-parameter settings!
- Performance gains are **mostly on easy-to-read and frequent-word inputs** – less than one quarter of the CNN/Dailynews data and practically no gains on RedditTIFU.
- BART-R3F **lacks robustness against training data variability** – new random split on RedditTIFU negates gains reported for split used in paper.

Enjoy reading the paper!

Data, code and additional material:

https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods/

Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., and Gupta, S. (2021).
Better fine-tuning by reducing representational collapse.
In *International Conference on Learning Representations (ICLR)*.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015).
Teaching machines to read and comprehend.
In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada.

Kim, B., Kim, H., and Kim, G. (2019).
Abstractive summarization of Reddit posts with multi-level memory networks.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, Minneapolis, Minnesota.

Kincaid, J. P., Fishburn, R. P., Rogers, R. L., and Chissom, B. S. (1975).
Derivation of new readability formulas for navy enlisted personnel.
Technical report, Technical Report, Naval Air Station, Millington, TN.

Lin, C.-Y. and Hovy, E. (2003).
Automatic evaluation of summaries using n-gram co-occurrence statistics.
In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada.

Platanios, E. A., Stretcu, O., Neubig, G., Poczos, B., and Mitchell, T. (2019).

Competence-based curriculum learning for neural machine translation.
In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*, Minneapolis, Minnesota.

Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., and Huang, X. (2020).
Extractive summarization as text matching.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online.