Recap: Inferential Reproducibility - A Worked-Through Example

BART-RXF: Better Fine-Tuning by Reducing Representational Collapse [Aghajanyan et al., 2021]

- SOTA on `paperswithcode.com` for text summarization task on CNN/Dailymail and RedditTIFU datasets.
- Baseline: BART [Lewis et al., 2019]
- SOTA Model: Approximate trust region method by constraining updates on embeddings f and classifier g during fine-tuning in order not to forget original pre-trained representations.

$$\mathcal{L}_{R3F}(f, g, \theta) = \mathcal{L}(\theta) + \lambda KL(g \cdot f(x) || g \cdot f(x + z))$$
$$\text{s.t. } z \sim \mathcal{N}(0, \sigma^2 I) \text{ or } z \sim \mathcal{U}(-\sigma, \sigma).$$

Experimental setup and SOTA results

- Datasets hosted on `paperwithcode.com`
  - train/dev/test split for Reddit not given, used split of [Zhong et al., 2020] instead.
- Reported meta-parameter ranges: $\lambda \in [0.001, 0.01, 0.1]$, noise distribution $\mathcal{N}$ or $\mathcal{U}$, maximum result of 10 random seeds .
  - Seeds of random number generator not given, used new 18 random seeds for baseline and 5 for SOTA.
- Results reported in [Aghajanyan et al., 2021]:

	CNN/DailyMail	Gigaword	Reddit TIFU (Long)
Random Transformer	38.27/15.03/35.48	35.70/16.75/32.83	15.89/1.94/12.22
BART	44.16/21.28/40.90	39.29/20.09/35.65	24.19/8.12/21.31
PEGASUS	44.17/**21.47**/41.11	39.12/19.86/36.24	26.63/9.01/21.60
ProphetNet (Old SOTA)	44.20/21.17/**41.30**	39.51/20.42/**36.69**	-
BART+R3F (New SOTA)	**44.38/21.53/41.17**	**40.45/20.69/36.56**	**30.31/10.98/24.74**

Significance Testing for Training Reproducibility

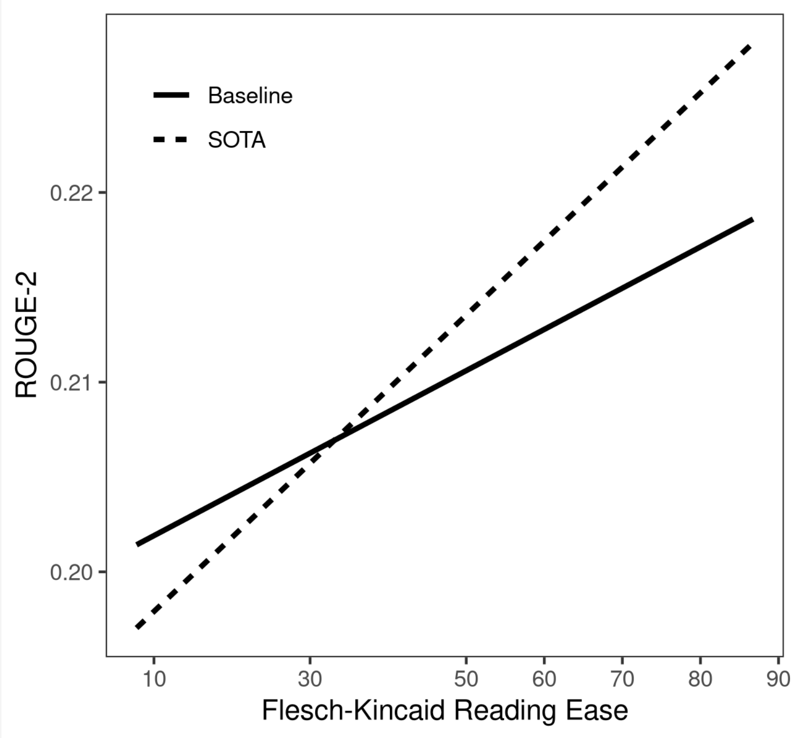
baseline - SOTA	p-value	effect size
Rouge1	$1.99e-14$	-0.101
Rouge2	0.00000000114	-0.0803
RougeL	$1.35e-15$	-0.105

- Rouge [Lin and Hovy, 2003] evaluation of best baseline versus best SOTA model on CNN/DailyMail shows **significant improvements of best SOTA model over baseline** with small effect sizes.

Measuring difficulty of summarization data

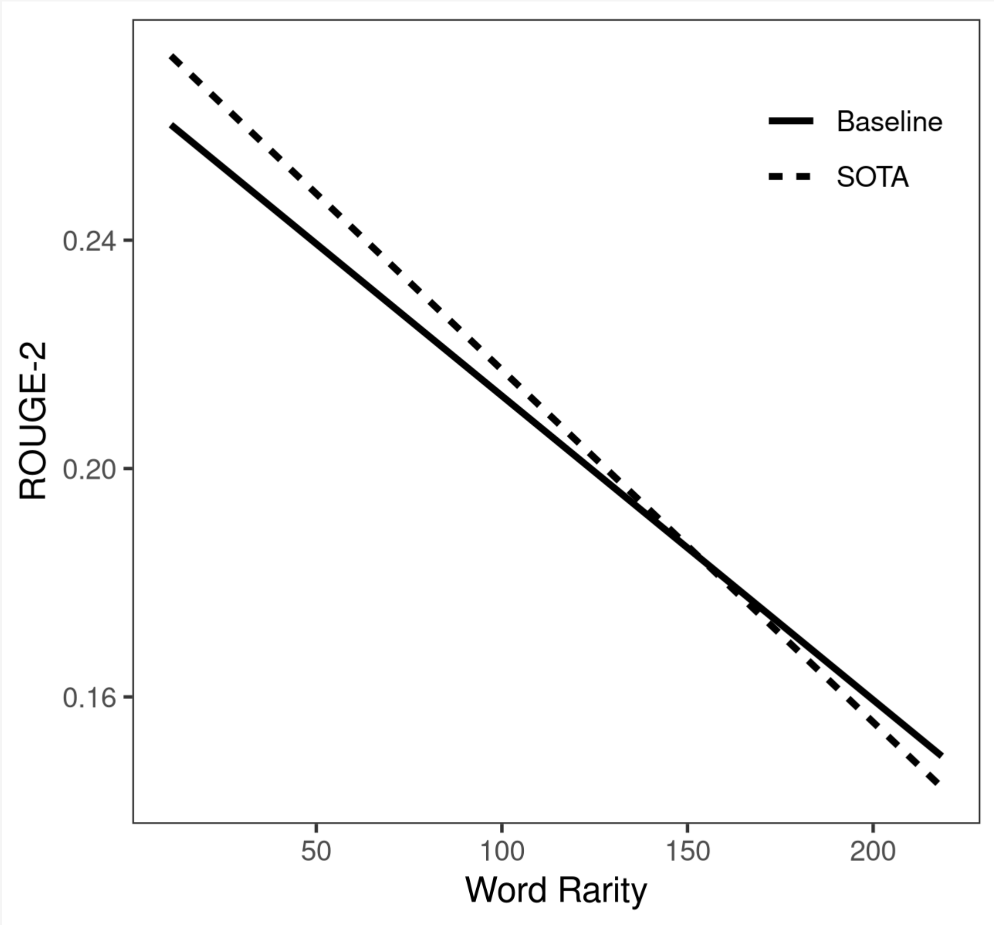
- **Word rarity** [Platanios et al., 2019]: Negative log of empirical probabilities of words in segment, higher value means higher rarity.
- **Flesch-Kincaid readability** [Kincaid et al., 1975]: Pro-rates words/sentences and syllables/word; in principle unbounded, but interpretation scheme exists for ranges from 0 (difficult) to 100 (easy).

Interaction of Performance with Data Properties



- Significant difference in performance slope with respect to ease of readability.
- Performance for SOTA system increases faster for easier inputs than for baseline.

Interaction of Performance with Data Properties



- Significant difference in performance with respect to word rarity.
- SOTA is better than baseline for inputs with lower word rarity.

Incorporating meta-parameter variation into significance testing

- Grid search over 18 random seeds for baseline, 30 SOTA models for 3 λ values $\times$ 2 noise distributions $\times$ 5 random seeds.

baseline - SOTA	p-value	effect size
Rouge1	0.0	0.390
Rouge2	0.0	0.301
RougeL	0.0	0.531

- **Relations turned around: Baseline significantly better than SOTA**, at medium effect size!
- Performance variation of baseline model over 18 random seeds negligible (standard deviations $< 0.2\%$ for Rouge-X scores)
- ➡ Reliability analysis of SOTA model!

Reliability coefficient and variance component analysis

Variance component v	Variance σ_v^2	Percent
summary_id	0.00923	55.7
lambda	0.00254	15.3
random_seed	0.000122	0.73
noise_distribution	0.0000473	0.29
residual	0.00464	28.0

- Only moderate value of reliability coefficient.
- Largest variance component for Rouge1 estimate due to regularization constant λ.

Reliability coefficient and variance component analysis

Variance component v	Variance σ_v^2	Percent
summary_id	0.00992	62.7
lambda	0.00131	8.31
random_seed	0.0000766	0.48
noise_distribution	0.0000318	0.2
residual	0.00449	28.3

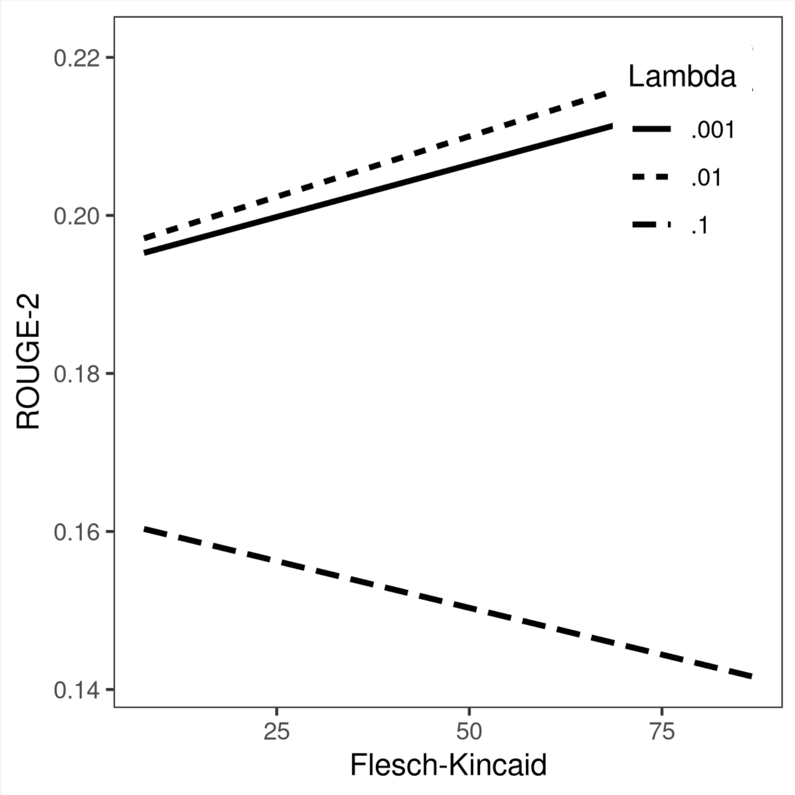
- Only moderate value of reliability coefficient.
- Largest variance component for Rouge2 estimate due to regularization constant λ.

Reliability coefficient and variance component analysis

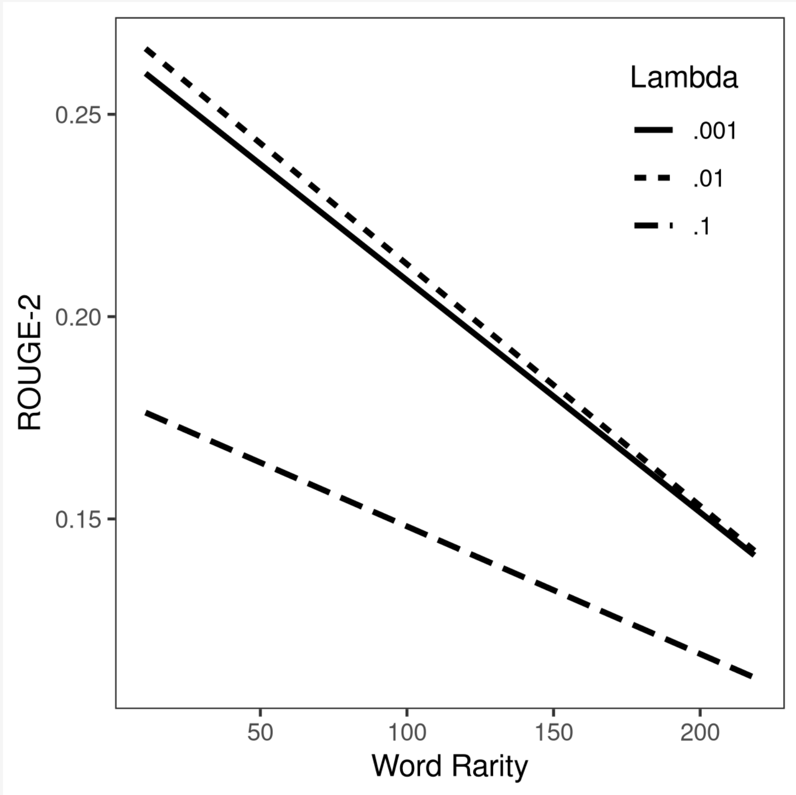
Variance component v	Variance σ_v^2	Percent
summary_id	0.00875	47.9
lambda	0.00519	28.4
random_seed	0.0000370	0.2
noise_distribution	0.0000144	0.08
residual	0.00428	23.4

- Poor value of reliability coefficient.
- Largest variance component for RougeL estimate due to regularization constant λ.

Interaction of Meta-Parameters with Data Properties



- Significant drop in performance of SOTA model across levels of reading difficulty for regularization constant $\lambda = 0.1$.



- Significant drop in performance of SOTA model for regularization constant $\lambda = 0.1$, especially for rare words.

Reproducibility Results on RedditTIFU

- Interesting data since much harder to read (mean readability score of -348.9).
- Significant improvement of best SOTA over baseline only for Rouge2 at small effect size.
- No significant improvements of SOTA over baseline if meta-parameter variation is taken into account.
- Reliability coefficients of around 80% with negligible variance contributions from λ values.

- Losing or winning a new SOTA score 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 improvements by fine-tuning **mostly on easy-to-read and frequent-word inputs** – less than one quarter of the CNN/Dailynews data.
- **Lacking robustness against data variability** – new random split on RedditTIFU negates gains reported for split used in paper.