Reliability and Learnability of Human Bandit Feedback for Sequence-to-Sequence Reinforcement Learning

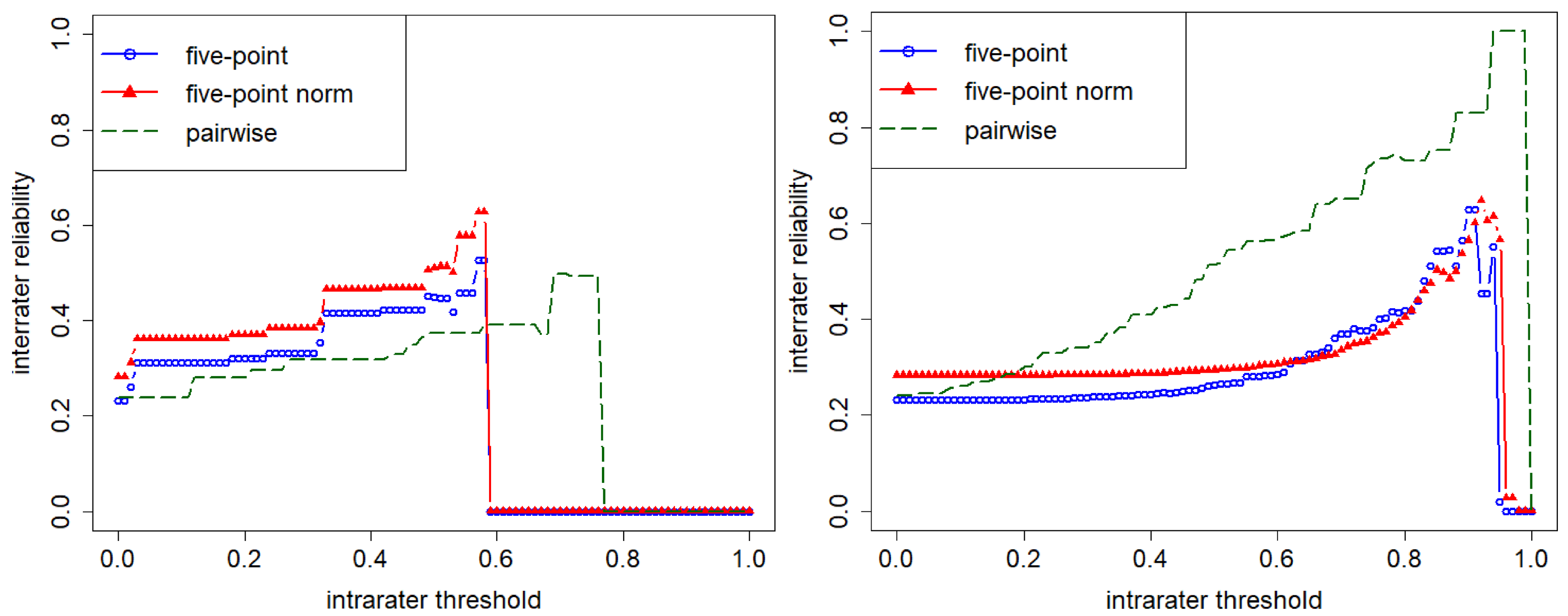
Julia Kreutzer[1] Joshua Uyheng[3] Stefan Riezler[1,2]

[1]CL & [2]IWR, Heidelberg University, Germany, [3]Deps. of Psychology & Mathematics, Ateneo de Manila University, Philippines

1. Reliability

	Inter-rater	Intra-rater	
Rating Type	α	Mean α	Stdev α
5-point	0.2308	0.4014	0.1907
+ normalization	0.2820		
+ filtering	**0.5059**	0.5527	0.0470
Pairwise	0.2385	0.5085	0.2096
+ filtering	0.3912	**0.7264**	0.0533

Table 1: Measuring **inter- and intra-rater reliability** with Krippendorff's α.



(a) Filtering by rater variance. (b) Filtering by item variance.

Figure 1: **Ablation analysis** on rater and item variance by filtering.

Rating Type	Avg. subjective difficulty [1-10]
5-point	4.8
Pairwise	5.69

Table 2: Subjective difficulty, judged by raters.

Difficulties with **5-point** ratings:
- Weighing of error types; long sentences with few essential errors

Difficulties with **Pairwise** ratings (incl. ties):
- Distinction between similar or similarly bad translations
- No normalization for individual biases
- Ties: no absolute anchoring of the quality of the pair

Summary

*Are **pairwise** ratings better for human reinforcement learning in NMT than standard 5-point ratings?*

- Collected & analyzed ~15 ratings for **800 translations**.
- Both have **comparable** inter-/intra-annotator α-agreement.
- Up to 1.1 BLEU improvement with **reward estimator**

⇒ Best reliability, learnability, and NMT gains for normalized, filtered **5-point** feedback.

Data: `http://www.cl.uni-heidelberg.de/statnlpgroup/humanmt/`

2. Learnability

Model	Feedback	Spearman's ρ
MSE	5-point norm.	-0.2193
	+ filtering	**-0.2341**
PW	Pairwise	-0.1310
	+ filtering	-0.1255

Table 3: Correlation between estimated rewards and TER.

Overcome feedback sparsity with a **reward estimator** $\hat{r}_{\psi}(\cdot)$.

5-point feedback: standard MSE on scaled ratings.

$$\mathcal{L}^{MSE}(\psi) = \frac{1}{n}\sum_{i=1}^{n}(r(\mathbf{y}_i) - \hat{r}_{\psi}(\mathbf{y}_i))^2.$$

Pairwise feedback: predict human preferences $Q[\cdot \succ \cdot]$.

$$\mathcal{L}^{PW}(\psi) = -\frac{1}{n}\sum_{i=1}^{n} Q[\mathbf{y}_i^1 \succ \mathbf{y}_i^2] \log \hat{P}_{\psi}[\mathbf{y}_i^1 \succ \mathbf{y}_i^2] + Q[\mathbf{y}_i^2 \succ \mathbf{y}_i^1] \log \hat{P}_{\psi}[\mathbf{y}_i^2 \succ \mathbf{y}_i^1],$$

with the Bradley-Terry model for preferences

$$\hat{P}_{\psi}[\mathbf{y}^1 \succ \mathbf{y}^2] = \frac{\exp \hat{r}_{\psi}(\mathbf{y}^1)}{\exp \hat{r}_{\psi}(\mathbf{y}^1) + \exp \hat{r}_{\psi}(\mathbf{y}^2)}.$$

3. Reinforcement Learning

Model	Rewards	BLEU	METEOR	BEER
Baseline	-	27.0	30.7	59.48
OPL	5-point norm.	27.5	30.9	59.72
RL	5-point norm.	**28.1**	31.5	60.21
	+ filtering	**28.1**	**31.6**	**60.29**
RL	Pairwise	27.8	31.3	59.88

Table 4: **NMT** domain adaptation (WMT→TED) with offline human feedback.

Neural Machine Translation. Standard 1-layer Encoder-Decoder with MLP-Attention, pre-trained on 5.9M WMT17 translations from German to English. Training is continued with weak feedback only.

Off-Policy Learning (OPL) from Direct Rewards. Improve the MT system from a log $L = \{(\mathbf{x}^{(h)}, \mathbf{y}^{(h)}, r(\mathbf{y}^{(h)}))\}_{h=1}^{H}$ of rewarded translations from the logging system.

$$\mathcal{R}^{OPL}(\boldsymbol{\theta}) = \frac{1}{H}\sum_{h=1}^{H} r(\mathbf{y}^{(h)})\, \bar{p}_{\boldsymbol{\theta}}(\mathbf{y}^{(h)}|\mathbf{x}^{(h)}),$$

- Reweighting over mini-batch B: $\bar{p}_{\boldsymbol{\theta}}(\mathbf{y}^{(h)}|\mathbf{x}^{(h)}) = \frac{p_{\boldsymbol{\theta}}(\mathbf{y}^{(h)}|\mathbf{x}^{(h)})}{\sum_{b=1}^{B} p_{\boldsymbol{\theta}}(\mathbf{y}^{(b)}|\mathbf{x}^{(b)})}$
- Only logged translations are reinforced, i.e. no exploration

RL from Estimated Rewards. Expected estimated reward maximization (REINFORCE), approximated with k samples (→MRT):

$$\mathcal{R}^{RL}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x})p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x})}\left[\hat{r}_{\psi}(\mathbf{y})\right] \approx \sum_{s=1}^{S}\sum_{i=1}^{k} p_{\boldsymbol{\theta}}^{\tau}(\tilde{\mathbf{y}}_i^{(s)}|\mathbf{x}^{(s)})\, \hat{r}_{\psi}(\tilde{\mathbf{y}}_i)$$

- Softmax temperature controls sharpness of sampling distribution $p_{\boldsymbol{\theta}}^{\tau}(\mathbf{y}|\mathbf{x}) = \text{softmax}(\mathbf{o}/\tau)$, i.e. the amount of exploration
- Subtract the running average of rewards from $\hat{r}_{\psi}$ to reduce variance (baseline control variate)