Improving Machine Translation with Human Feedback: An Exploration of Quality Estimation as a Reward Model

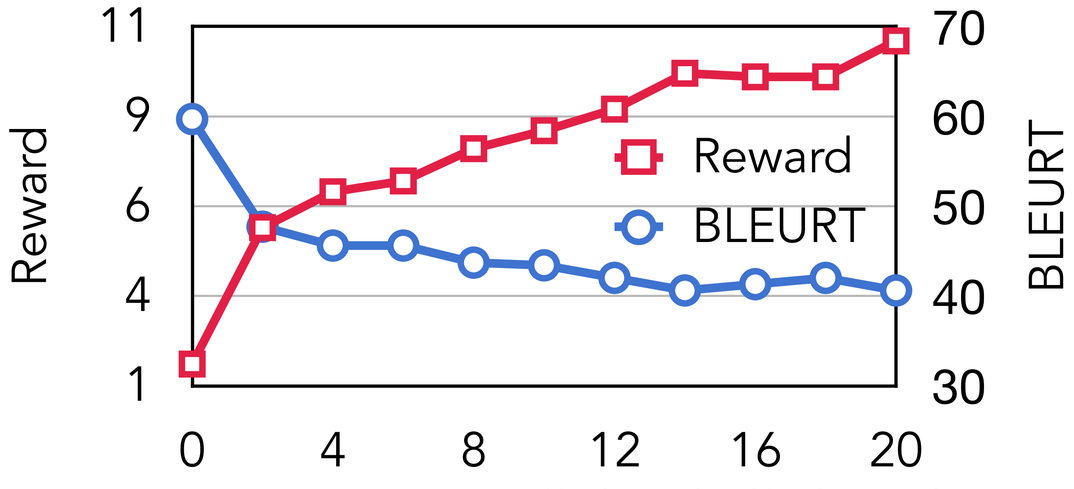
Zhiwei He, Xing Wang*, Wenxiang Jiao, Zhuosheng Zhang,
Rui Wang*, Shuming Shi, Zhaopeng Tu

Check out our paper for more details ☝🏻

Why QE as a reward model?

- ✔ LLMs benefit from human preference modeled by reward models.
- ✔ Today's QE models (reference-free) closely match human preference.
- ? Can MT model learning from QE?

Direct use: over-optimization (OO)



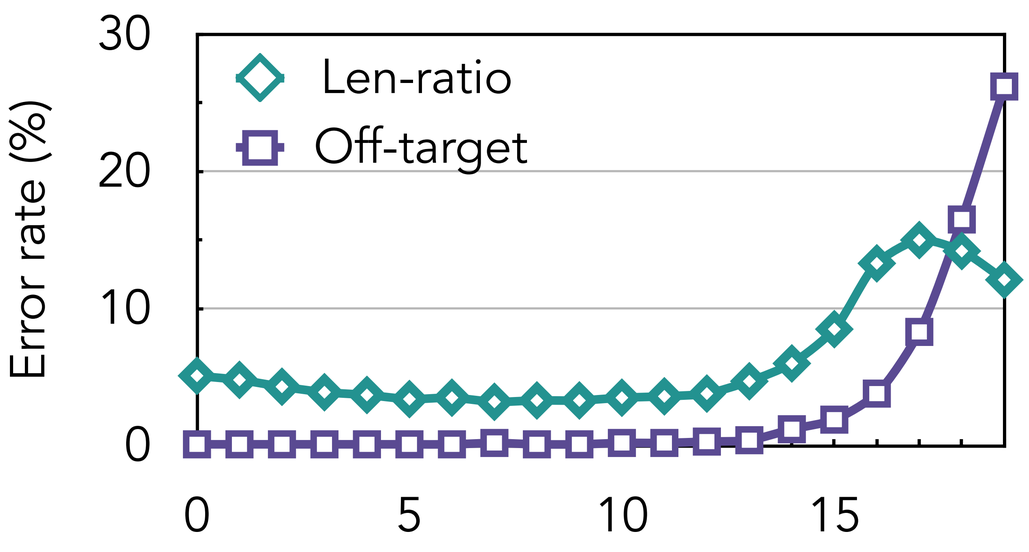
- Using RAFT as an off-the-shelf algorithm.
- Overoptimizing rewards could steer the model away from human preferences.

OO Cause 1: imperfect reward

Error type	Translation	Reward
None	The rule of drinking Red Label Whisky:	2.84
Len-ratio (too long/short translation)	The rule of drinking Red Label Whisky: 1. Always drink responsibly.2. Never drink alone.3. Avoid drinking on an empty stomach.4. Set limits and stick to them.5. Drink in moderation.	5.60
Off-target (wrong target language)	So trinkt man Red-Label-Whisky:	4.58

▸ *QE sometimes assigns high scores for errors.*

OO Cause 2: error propagation

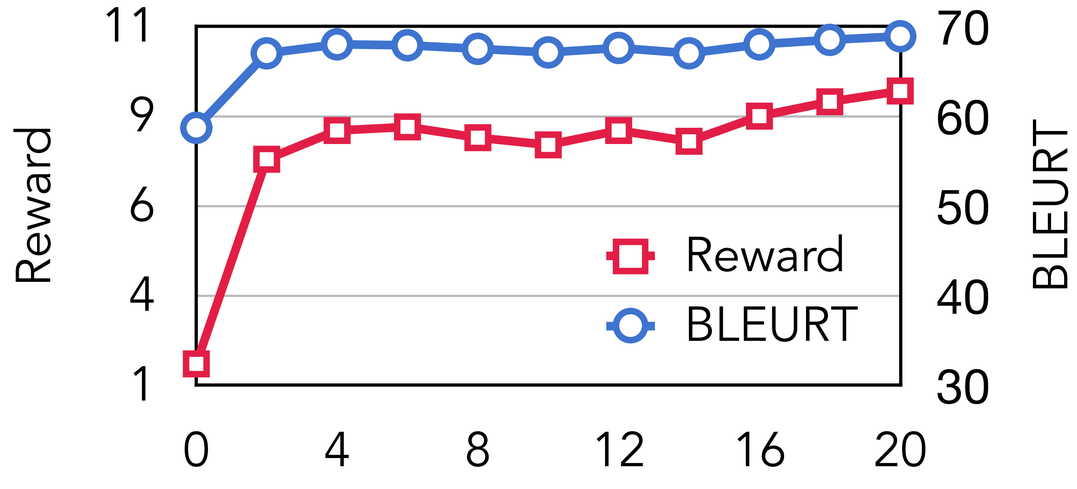


▸ *Errors rapidly propagate in training.*

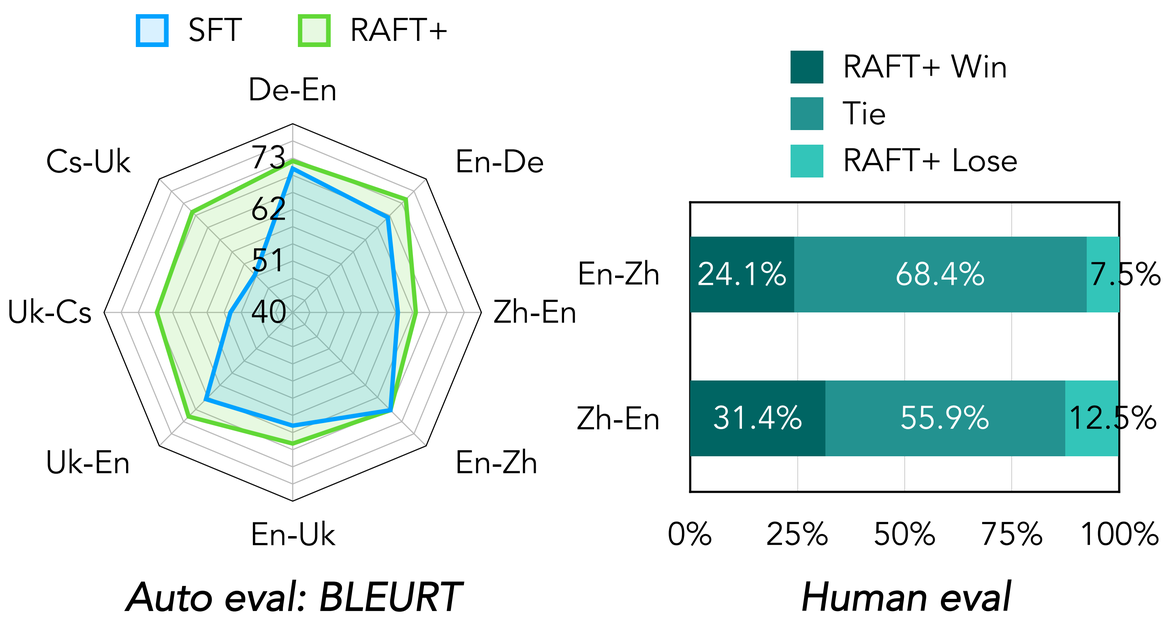
Mitigating OO

$$r^+(x,y) = \begin{cases} r(x,y) - P & \text{if } C(x,y) \\ r(x,y) & \text{otherwise,} \end{cases}$$

- Detect errors during training
- Add penalty term P to the reward if y is an error translation (C(x, y) = True)
- We dub it as RAFT+

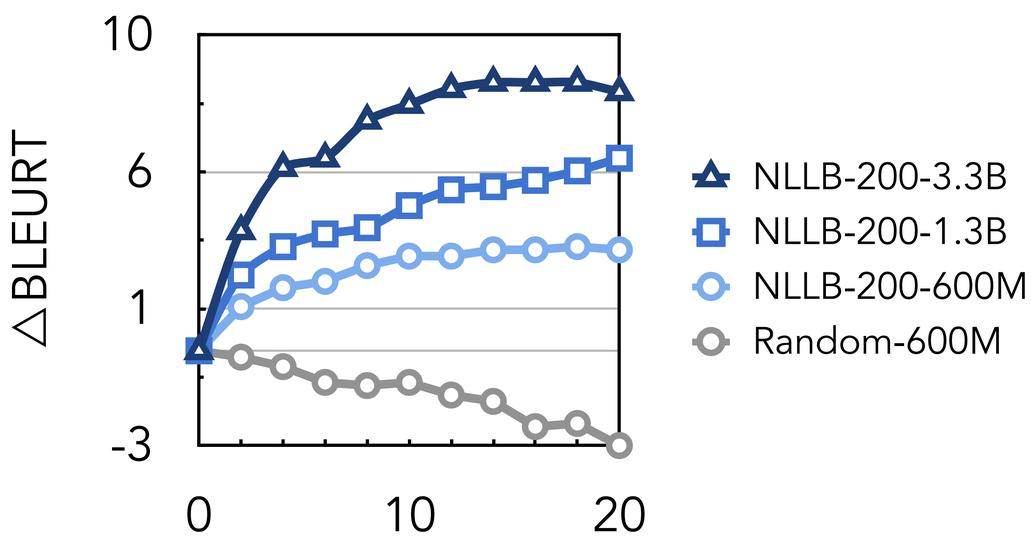


Training curve



Auto eval: BLEURT *Human eval*

Effects of base model



- Pipeline: Base -> SFT -> RAFT+
- Larger model size results in a more significant enhancement from RAFT+.
- RAFT+ only works when the base model has undergone pretraining.