Reliability & Learnability of Human Bandit Feedback for Seq2Seq Reinforcement Learning

Julia Kreutzer
Heidelberg University, Germany

Joint work with Joshua Uyheng (Ateneo de Manila University) & Stefan Riezler (Heidelberg University)

Lost in Translation

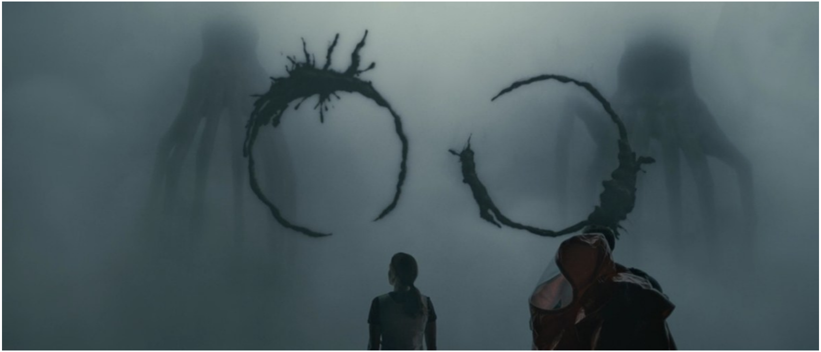


Figure 1: Arrival. https://glyphpress.com/talk/2017/the-journey-is-the-arrival

Machine Translation

	Source x	Target y
Training		*"it wasn't us"*
		*"there is no linear time"*
Testing		?

Supervised training with a parallel corpus $\mathcal{D}$ of sources & targets:

$$\mathcal{L}^{\text{MLE}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \log p_\theta(y^{(i)} \mid x^{(i)})$$

$$= \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sum_{t=1}^{T_y} \log p_\theta(y_t^{(i)} \mid x_t^{(i)}, y_{<t}^{(i)})$$

Machine Translation

	Source x	Target y
Training		*“it wasn’t us”*
		*“there is no linear time”*
Testing		*“life choice”* $\leftrightarrow$ *“you have to choose life”*

Supervised training with a parallel corpus $\mathcal{D}$ of sources & targets:

$$\mathcal{L}^{\text{MLE}}(\theta) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\log p_\theta(y^{(i)} \mid x^{(i)}) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\sum_{t=1}^{T_y}\log p_\theta(y_t^{(i)} \mid x_t^{(i)}, y_{<t}^{(i)})$$

Automatic evaluation: measure overlap with reference translation(s).

Neural Machine Translation

Neural Machine Translation (NMT) as seq2seq task with challenges in

1. Deep Learning: Train Encoder-Decoder architectures.
    - Structured outputs with long-range dependencies
    - Data sparsity and noise
    - Linguistic interpretability

Neural Machine Translation

Neural Machine Translation (NMT) as seq2seq task with challenges in

1. Deep Learning: Train Encoder-Decoder architectures.
   - Structured outputs with long-range dependencies
   - Data sparsity and noise
   - Linguistic interpretability
2. Reinforcement Learning: Maximize expected reward.
   - Large discrete action space
   - Underspecified reward functions
   - Sparse rewards

Learning from Humans

Improving NMT with **human bandit feedback**

- Cheaper than references
- No experts required
- Ideal for interactive usecases
- Fast model adaptation

Learning from Humans

Improving NMT with **human bandit feedback**

- Cheaper than references
- No experts required
- Ideal for interactive usecases
- Fast model adaptation

Challenges

- **Humans**: biased judgment and variance
- **Machine**: needs exploration, data-hungry

Improving MT with Weak Feedback

Learning from simulated...

- Online Bandit Feedback:
  - REINFORCE for SMT & NMT (Sokolov et al., 2016; Kreutzer et al., 2017)
  - Advantage Actor Critic for NMT (Nguyen et al., 2017; Lam et al., 2018)
  - WMT shared task: Amazon product titles (Sokolov et al., 2017)
- Offline Bandit Feedback:
  - Counterfactual learning for SMT (Lawrence et al., 2017b,a)

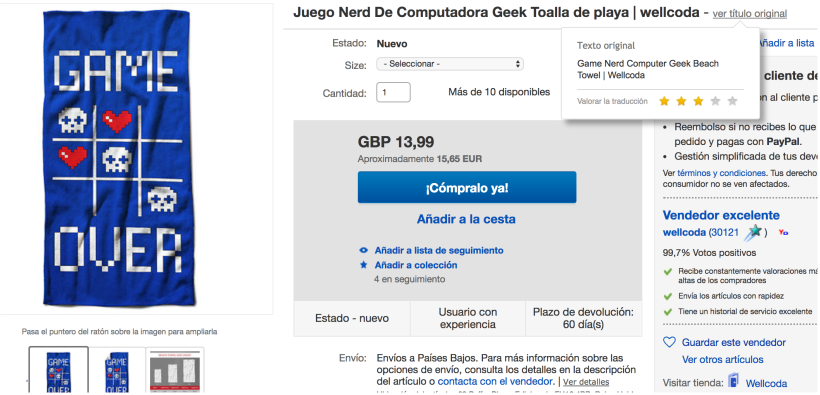
Improving MT with Weak Feedback

Learning from simulated...

- Online Bandit Feedback:
  - REINFORCE for SMT & NMT (Sokolov et al., 2016; Kreutzer et al., 2017)
  - Advantage Actor Critic for NMT (Nguyen et al., 2017; Lam et al., 2018)
  - WMT shared task: Amazon product titles (Sokolov et al., 2017)
- Offline Bandit Feedback:
  - Counterfactual learning for SMT (Lawrence et al., 2017b,a)

Today: Improve NMT with offline bandit feedback from **humans**.

No Success with Explicit User Feedback (Kreutzer et al., 2018a)



☹ Learning from 70k eBay user ratings fails due to **unreliable** ratings.

Success with Implicit Feedback (Kreutzer et al., 2018a)

Embed the feedback collection into a “back-translation” CLIR task:

query (es) $\xrightarrow[\text{translation}]{\text{query}}$ query (en) $\xrightarrow{\text{search}}$ title (en) $\xrightarrow[\text{translation}]{\text{item}}$ title (es)

“candado bicicleta” → “bicycle lock” → “...lock bike” → “...cerradura bicicleta”

Success with Implicit Feedback (Kreutzer et al., 2018a)

Embed the feedback collection into a "back-translation" CLIR task:

$$\text{query (es)} \xrightarrow[\text{translation}]{\text{query}} \text{query (en)} \xrightarrow{\text{search}} \text{title (en)} \xrightarrow[\text{translation}]{\text{item}} \text{title (es)}$$

"candado bicicleta" → *"bicycle lock"* → *"...lock bike"* → *"...cerradura bicicleta"*

⇒ **Task-specific reward function**: translated words match the query.

☺ Translation improves significantly!

Pairwise Preferences to the Rescue?

Does Thurstone (1927)'s **law of comparative judgment** hold for MT?

Source: *"Sie gehen im Geiste durch dieses Haus, in demn Sie wohnen, und schauen sich an, wie viele Türen da sind."*

NMT_1: *"They go in the spirit through this house, in the back of them, and look at how many doors there are."*
NMT_2: *"You go in the spirit of this house, in demn you live, and look at how many doors are there."*

Target: *"In your mind, you are walking through the house where you live, and are seeing how many doors there are."*

Controlled Feedback Collection (Kreutzer et al., 2018a)

TRANSLATION: Now i'm saying, "computer, take the 10 percent of the sequences that have come to my prescription. *
ORIGINAL: Jetzt sage ich, "Computer, nimm jetzt diejenigen 10 % der Sequenzen, welche meinen Vorgaben am nächsten gekommen sind.

- 5 (Very Good)
- 4 (Good)
- 3 (Neither Good nor Bad)
- 2 (Bad)
- 1 (Very Bad)

VS

ORIGINAL: Der andere Hut, den ich bei meiner Arbeit getragen habe, ist der der Aktivistin, als PatientInnenanwältin -- oder, wie ich manchmal sage, als ungeduldige Anwältin -- von Menschen, die Patienten von Ärzten sind. *

- TRANSLATION 1: The other hat i worn at my work is the activist, as a patient woman -- or, as i sometimes say, as an impatient lawyer -- of people who are patients of doctors.
- TRANSLATION 2: The other hat i've carried in my work is the activist, the patient's lawyer -- or, as i say sometimes, as an impatient lawyer -- of people who are patients of doctors.
- NO PREFERENCE

Collected feedback from ~15 bilinguals for 800 translations

1. **Reliability**: How reliable is each type of feedback?
2. **Learnability**: How well can we model this feedback?
3. **RL**: How much can it improve our NMT model?

Reliability

	Inter-rater	Intra-rater	
Rating Type	α	Mean α	Stdev α
5-point	0.2308	0.4014	0.1907
Pairwise	0.2385	0.5085	0.2096

Reliability

Rating Type	Inter-rater α	Intra-rater Mean α	Intra-rater Stdev α
5-point	0.2308	0.4014	0.1907
+ normalization	0.2820		
Pairwise	0.2385	0.5085	0.2096

Reliability

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

$\Rightarrow$ Pairwise ratings turn out to be **more difficult**.

Learnability: Reward Estimators

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

1. Tackle the arguably simpler problem of learning a reward estimator from human feedback first.
2. Then provide unlimited learned feedback to generalize to unseen outputs in off-policy RL.

End-to-End RL: Direct Rewards

Off-Policy Learning (OPL) from Direct Rewards

Improve the target NMT system (θ) with logged rewarded translations of the deterministic logging system. (Lawrence et al., 2017b)

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

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

End-to-End RL: Estimated Rewards

RL from Estimated Rewards

Reinforce k translation samples for each input with estimated rewards $\hat{r}_\psi$ for an approximation of the expected estimated reward.

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

- Similar to minimum risk training for NMT (Shen et al., 2016)
- Softmax temperature τ to control the amount of exploration
- Subtract the running average of rewards from $\hat{r}_\psi$ to reduce gradient variance and estimation bias.

Results on TED Talk Translations

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

- OPL uses 800 human rewards directly ⇒ overfitting
- RL (or MRT) uses unlimited amount of estimated rewards

Summary: Deep RL from Human Feedback Signals for NMT

1. Experiments with eBay product title translations (Kreutzer et al., 2018a)
    - Failed with **explicit** 5-star user ratings on a **large** collection of product title translations — feedback too noisy
    - Succeeded with **implicit** task-based feedback collected in a cross-lingual search task — well-defined reward function

Summary: Deep RL from Human Feedback Signals for NMT

1. Experiments with eBay product title translations (Kreutzer et al., 2018a)
    - Failed with **explicit** 5-star user ratings on a **large** collection of product title translations — feedback too noisy
    - Succeeded with **implicit** task-based feedback collected in a cross-lingual search task — well-defined reward function
2. Reliability and Learnability of Human Reinforcement (Kreutzer et al., 2018b)
    - Influence of reliability of 5-point ratings and pairwise preferences
    - Success with **explicit** 5-point ratings on a **small** set of TED talk translations — controlled feedback collection

Summary: Deep RL from Human Feedback Signals for NMT

1. Experiments with eBay product title translations (Kreutzer et al., 2018a)
    - Failed with **explicit** 5-star user ratings on a **large** collection of product title translations — feedback too noisy
    - Succeeded with **implicit** task-based feedback collected in a cross-lingual search task — well-defined reward function
2. Reliability and Learnability of Human Reinforcement (Kreutzer et al., 2018b)
    - Influence of reliability of 5-point ratings and pairwise preferences
    - Success with **explicit** 5-point ratings on a **small** set of TED talk translations — controlled feedback collection

Recipe?

- Reduce human biases and difficulties during feedback collection
- Encode human domain knowledge in learned reward estimator
- Use learned reward function as feedback signal in RL

Thank you!

Questions?

kreutzer@cl.uni-heidelberg.de

References

References I

Kreutzer, J., Khadivi, S., Matusov, E., and Riezler, S. (2018a). Can neural machine translation be improved with user feedback? In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Industry Track (NAACL-HLT)*, New Orleans, LA.

Kreutzer, J., Sokolov, A., and Riezler, S. (2017). Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada.

References II

Kreutzer, J., Uyheng, J., and Riezler, S. (2018b). Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Lam, T. K., Kreutzer, J., and Riezler, S. (2018). A reinforcement learning approach to interactive-predictive neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT)*, Alicante, Spain.

Lawrence, C., Gajane, P., and Riezler, S. (2017a). Counterfactual learning for machine translation: Degeneracies and solutions. In *Proceedings of the NIPS WhatIF Workshop*, Long Beach, CA.

References III

Lawrence, C., Sokolov, A., and Riezler, S. (2017b). Counterfactual learning from bandit feedback under deterministic logging : A case study in statistical machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Nguyen, K., Daumé III, H., and Boyd-Graber, J. (2017). Reinforcement learning for bandit neural machine translation with simulated human feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark.

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. (2016). Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Sokolov, A., Kreutzer, J., Riezler, S., and Lo, C. (2016). Stochastic structured prediction under bandit feedback. In *Advances in Neural Information Processing Systems*, Barcelona, Spain.

Sokolov, A., Kreutzer, J., Sunderland, K., Danchenko, P., Szymaniak, W., Fürstenau, H., and Riezler, S. (2017). A shared task on bandit learning for machine translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34:278–286.