Seq2seq RL for NMT: From Simulations to Human RL

- Where do simulations fall short?
  - Real-wold RL only has access to **human bandit feedback** $\Rightarrow$ **control variates**
  - Online/on-policy learning raises **safety and stability concerns** $\Rightarrow$ **offline learning**
  - **Human rewards** are **not well defined, noisy, and skewed** $\Rightarrow$ **reward estimation**

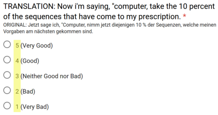
Offline Learning from Human Feedback: e-commerce

- [Kreutzer et al., 2018]: 69k translated item titles (en-es) with 148k individual ratings
- No agreement of paid raters with e-commerce users, low inter-rater agreement, learning impossible

Offline Learning from Human Feedback: e-commerce

- Lessons from e-commerce experiments:
  - Offline learning from direct user feedback to e-commerce titles is equivalent to **learning from noise**
  - Conjecture: Missing reliability and validity of human feedback in e-commerce experiment
  - Need experiment on controlled feedback collection!

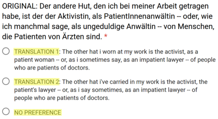
Offline Learning from Controlled Human Feedback



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

- Comparison of judgments on five-point Likert scale to pairwise preferences
- Feedback collected from ∼15 bilinguals for 800 translations (de-en)[1]

[1] Data: https://www.cl.uni-heidelberg.de/statnlpgroup/humanmt/

Reliability and Learnability of Human Feedback

- Controlled study on main factors in human RL:
    1. **Reliability**: Collect five-point and pairwise feedback on same data, evaluate intra- and inter-rater agreement.
    2. **Learnability**: Train reward estimators on human feedback, evaluate correlation to TER on held-out data.
    3. **RL**: Use rewards directly or estimated rewards to improve an NMT system.

What are your guesses on reliability and learnability—five-point or pairwise?

Reliability: α-agreement

Rating Type	Inter-rater α	Intra-rater Mean α	Intra-rater 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

- Inter- and intra-reliability measured by Krippendorff's α for 5-point and pairwise ratings of 1,000 translations of which 200 translations are repeated twice.
- Filtered variants are restricted to either a subset of participants or a subset of translations.

Reliability: Qualitative Analysis

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

- Difficulties with **5-point** ratings:
  - Weighing of error types; long sentences with few essential errors
- Difficulties with **Pairwise** ratings (incl. ties):
  - Distinction between similar translations
  - Ties: no absolute anchoring of the quality of the pair
  - Final score: No normalization for individual biases possible

Learnability: 5-point Feedback

- Inputs are sources $\mathbf{x}$ and their translations $\mathbf{y}$
- Given cardinal ratings r, train a regression model with parameters ψ to minimize the mean squared error (MSE) for predicted rewards $\hat{r}$:

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

Learnability: Pairwise Feedback

- Given human preference $Q[\mathbf{y^1} \succ \mathbf{y^2}]$ for translation $\mathbf{y_1}$ over translation $\mathbf{y_2}$
- Train estimator $\hat{P}_\psi[\mathbf{y^1} \succ \mathbf{y^2}]$ by minimizing cross-entropy between predictions and human preferences:

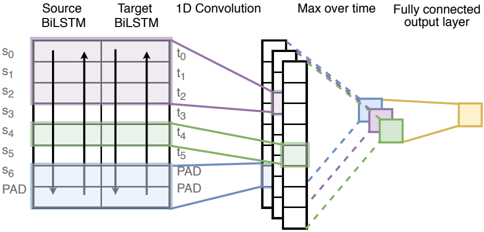
$$L(\psi) = -\frac{1}{n}\sum_{i=1}^{n}\Big(Q[\mathbf{y_i^1} \succ \mathbf{y_i^2}] \log \hat{P}_\psi[\mathbf{y_i^1} \succ \mathbf{y_i^2}] + Q[\mathbf{y_i^2} \succ \mathbf{y_i^1}] \log \hat{P}_\psi[\mathbf{y_i^2} \succ \mathbf{y_i^1}]\Big),$$

  with the Bradley-Terry model for preferences

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

- Use Bradley-Terry model's $\hat{r}$ as reward estimator [Christiano et al., 2017]

Reward Estimator Architecture



- biLSTM-enhanced bilingual extension of convolutional model for sentence classification [Kim, 2014]

Learnability: Results

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

- Comparatively better results for reward estimation from cardinal human judgements.
- Overall relatively low correlation, presumably due to overfitting on small training data set.

End-to-end Seq2seq RL

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 from Estimated Rewards

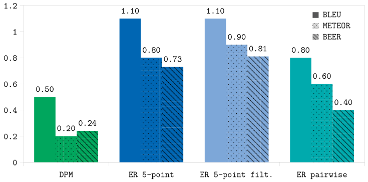
Expected Risk Minimiziation from Estimated Rewards

Estimated rewards allow to use minimum risk training [Shen et al., 2016] s.t. feedback can be collected for k samples:

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

- Softmax temperature τ to control the amount of exploration by sharpening the sampling distribution $p_\theta^\tau(\mathbf{y}|\mathbf{x}) = \mathrm{softmax}(\mathbf{o}/\tau)$ at lower temperatures.
- Subtract the running average of rewards from $\hat{r}_\psi$ to reduce gradient variance and estimation bias.

Results on TED Talk Translations



- Significant improvements over the baseline (27.0 BLEU / 30.7 METEOR / 59.48 BEER):
  - Gains of 1.1 BLEU for expected risk (ER) minimization for estimated rewards.
  - Deterministic propensity matching (DPM) on directly logged human feedback yields up to 0.5 BLEU points.

Summary

Basic RL:

- **Policy evaluation** using **Dynamic Programming**
- **Policy optimization** using **Dynamic Programming**, **Monte Carlo**, or both: **Temporal Difference** learning.
- **Policy-gradient** techniques for direct policy optimization.

Seq2seq RL:

- Seq2seq RL **simulations**: Bandit Neural Machine Translation.
- **Offline** learning from deterministically logged feedback: Deterministic Propensity Matching.
- Seq2seq RL from **human feedback**: Collecting reliable feedback, learning reward estimators, end-to-end RL from estimated rewards.

References

- Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., Courville, A., and Bengio, Y. (2017). An actor-critic algorithm for sequence prediction.
  In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France.

- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate.
  In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA.

- Bastani, H., Bayati, M., and Khosravi, K. (2017). Exploiting the natural exploration in contextual bandits.
  ArXiv e-prints, 1704.09011.

- Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: The example of computational advertising.
  Journal of Machine Learning Research, 14:3207–3260.

- Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
  Foundations and Trends in Machine Learning, 5(1):1–122.

- Chapelle, O., Masnavoglu, E., and Rosales, R. (2014). Simple and scalable response prediction for display advertising.
  ACM Transactions on Intelligent Systems and Technology, 5(4).
- Chen, M. X., Firat, O., Bapna, A., Johnson, M., Macherey, W., Foster, G., Jones, L., Schuster, M., Shazeer, N., Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Chen, Z., Wu, Y., and Hughes, M. (2018). The best of both worlds: Combining recent advances in neural machine translation.
  In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences.
  In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.
- Dudik, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning.
  In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, WA.
- Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. (2017). Convolutional sequence to sequence learning.
  In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, Vancouver, Canada.

- Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning.
  In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, New York, NY.
- Kim, Y. (2014). Convolutional neural networks for sentence classification.
  In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.
- Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms.
  In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada.
- Kong, A. (1992). A note on importance sampling using standardized weights.
  Technical Report 348, Department of Statistics, University of Chicago, Illinois.
- Kreutzer, J., Khadivi, S., Matusov, E., and Riezler, S. (2018). 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 (ACL)*, Vancouver, Canada.

- Langford, J., Strehl, A., and Wortman, J. (2008). Exploration scavenging.
  In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, Helsinki, Finland.
- 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.
- Lawrence, C. and Riezler, S. (2018). Improving a neural semantic parser by counterfactual learning from human bandit feedback.
  In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.
- 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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation.
  In *Proceedings of the 19th International Conference on World Wide Web (WWW)*.
- Nguyen, K., Daumé, H., and Boyd-Graber, J. (2017). Reinforcement learning for bandit neural machine translation with simulated feedback.

In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark.

- Precup, D., Sutton, R. S., and Singh, S. P. (2000). Eligibility traces for off-policy policy evaluation.
  In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, San Francisco, CA.
- Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference.
  In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Reykjavik, Iceland.
- Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2016). Sequence level training with recurrent neural networks.
  In *Proceedings of the International Conference on Learning Representation (ICLR)*, San Juan, Puerto Rico.
- Ross, S. M. (2013). *Simulation*.
  Elsevier, fifth edition.
- 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 (ACL)*, Berlin, Germany.

- Sokolov, A., Kreutzer, J., Lo, C., and Riezler, S. (2016). Stochastic structured prediction under bandit feedback.
  In *Advances in Neural Information Processing Systems (NIPS)*, Barcelona, Spain.
- Sokolov, A., Riezler, S., and Urvoy, T. (2015). Bandit structured prediction for learning from user feedback in statistical machine translation.
  In *Proceedings of MT Summit XV*, Miami, FL.
- Strehl, A. L., Langford, J., Li, L., and Kakade, S. M. (2010). Learning from logged implicit exploration data.
  In *Advances in Neural Information Processing Sytems (NIPS)*, Vancouver, Canada.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks.
  In *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada.
- Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning. An Introduction*.
  The MIT Press.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning. An Introduction*.
  The MIT Press, second edition.
- Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation.
  In *Advances in Neural Information Processings Systems (NIPS)*, Vancouver, Canada.

- Szepesvári, C. (2009). *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA.
- Watkins, C. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.