Bandit Structured Prediction for Neural Seq2Seq Learning

Julia Kreutzer, Artem Sokolov, Stefan Riezler

Heidelberg University, Germany

Bandit Structured Prediction

Algorithm 1 Bandit Structured Prediction

Input: Sequence of learning rates γ_k

Output: Optimal parameters $\hat{\theta}$

Initialize parameters θ_0

for $k = 0, \ldots, K$ **do**

  Observe input structure $\mathbf{x}_k$

  Sample output structure $\tilde{\mathbf{y}}_k \sim p_\theta(\mathbf{y}|\mathbf{x}_k)$

  Obtain feedback $\Delta(\tilde{\mathbf{y}}_k)$

  Compute stochastic gradient s_k

  Update parameters $\theta_{k+1} = \theta_k - \gamma_k \, s_k$

Choose a solution $\hat{\theta}$ from the list $\{\theta_0, \ldots, \theta_K\}$

Objectives

1 **Expected Loss (EL)**:

Expectation of a task loss $\Delta(\tilde{\mathbf{y}})$ over all input and output structures:

$$L^{\mathsf{EL}}(\theta) = \mathbb{E}_{p(\mathbf{x})\, p_\theta(\tilde{\mathbf{y}}|\mathbf{x})} \left[\Delta(\tilde{\mathbf{y}})\right].$$

Stochastic gradient:

$$s_k^{\mathsf{EL}} = \Delta(\tilde{\mathbf{y}}) \frac{\partial \log p_\theta(\tilde{\mathbf{y}}|\mathbf{x}_k)}{\partial \theta}$$

Output structures $\tilde{\mathbf{y}}$ are sampled **word by word** from the distribution resulting from the softmax transformation in the output layer of the network.

2 **Pairwise Preference Ranking (PR)**:

Transfer EL to **pairs of structures** $\langle \tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j \rangle$:

$$L^{\mathsf{PR}}(\theta) = \mathbb{E}_{p(\mathbf{x})\, p_\theta(\langle \tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j \rangle|\mathbf{x})} \left[\Delta(\langle \tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j \rangle)\right].$$

Stochastic gradient:

$$s_k^{\mathsf{PR}} = \Delta(\langle \tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j \rangle) \times \left(\frac{\partial \log p_\theta(\tilde{\mathbf{y}}_i|\mathbf{x}_k)}{\partial \theta} + \frac{\partial \log p_\theta^-(\tilde{\mathbf{y}}_j|\mathbf{x}_k)}{\partial \theta} \right).$$

Learn to rank $\tilde{\mathbf{y}}_i$ over $\tilde{\mathbf{y}}_j$ with **pairwise feedback**, either continuous (cont)

$$\Delta(\langle \mathbf{y}_i, \mathbf{y}_j \rangle) = \Delta(\mathbf{y}_j) - \Delta(\mathbf{y}_i),$$

or binary (bin)

$$\Delta(\langle \mathbf{y}_i, \mathbf{y}_j \rangle) = \begin{cases} 1 & \text{if } \Delta(\mathbf{y}_j) > \Delta(\mathbf{y}_i), \\ 0 & \text{otherwise.} \end{cases}$$

Draw **negative sample** $\tilde{\mathbf{y}}_j$ from distribution p_θ^-, one word per output structure (chosen randomly):

$$p_\theta^-(\tilde{y}_t = w_j|\mathbf{x}, \hat{\mathbf{y}}_{<t}) = \frac{\exp(-o_{w_j})}{\sum_{v=1}^{V} \exp(-o_{w_v})}.$$

Bandit Seq2Seq

Bandit structured prediction [1] is a stochastic optimization framework where learning is performed from **partial feedback**. This feedback is received in the form of task loss evaluation of a predicted output structure, without having access to gold standard structures.

In this work, we advance the framework by

- lifting linear bandits to **neural seq2seq learning** using attention-based RNNs, and
- incorporating **control variates** for variance reduction and improved generalization.

Experiments for **neural machine translation** show large improvements for domain adaptation from simulated bandit feedback.

Control Variates

Augment a random variable X (here: $X = s_k$) by another random variable Y, the control variate. With $\bar{Y} = \mathbb{E}[Y]$, $X - \hat{c}\,Y + \hat{c}\,\bar{Y}$ is an unbiased estimator of $\mathbb{E}[X]$. Control variates with high $\mathrm{Cov}(X, Y)$ **reduce the variance** of the gradient estimate. Two choices here:

1 **Baseline (BL)** [2]:

$$Y_k = \nabla \log p_\theta(\tilde{\mathbf{y}}|\mathbf{x}_k) \frac{1}{k} \sum_{j=1}^{k} \Delta(\tilde{\mathbf{y}}_j).$$

2 **Score Function (SF)** [3]:

$$Y_k = \nabla \log p_\theta(\tilde{\mathbf{y}}|\mathbf{x}_k).$$

Experiments

Neural machine translation **domain adaptation**:

- Adapt a pre-trained model (Europarl, fr-en) to new domains (News Commentary and TED).
- **Simulated feedback** with GLEU on references
- Encoder-decoder architecture with attention
- Full-information baselines: maximum likelihood estimation on reference translations

Strategies for **handling of unknown words**:

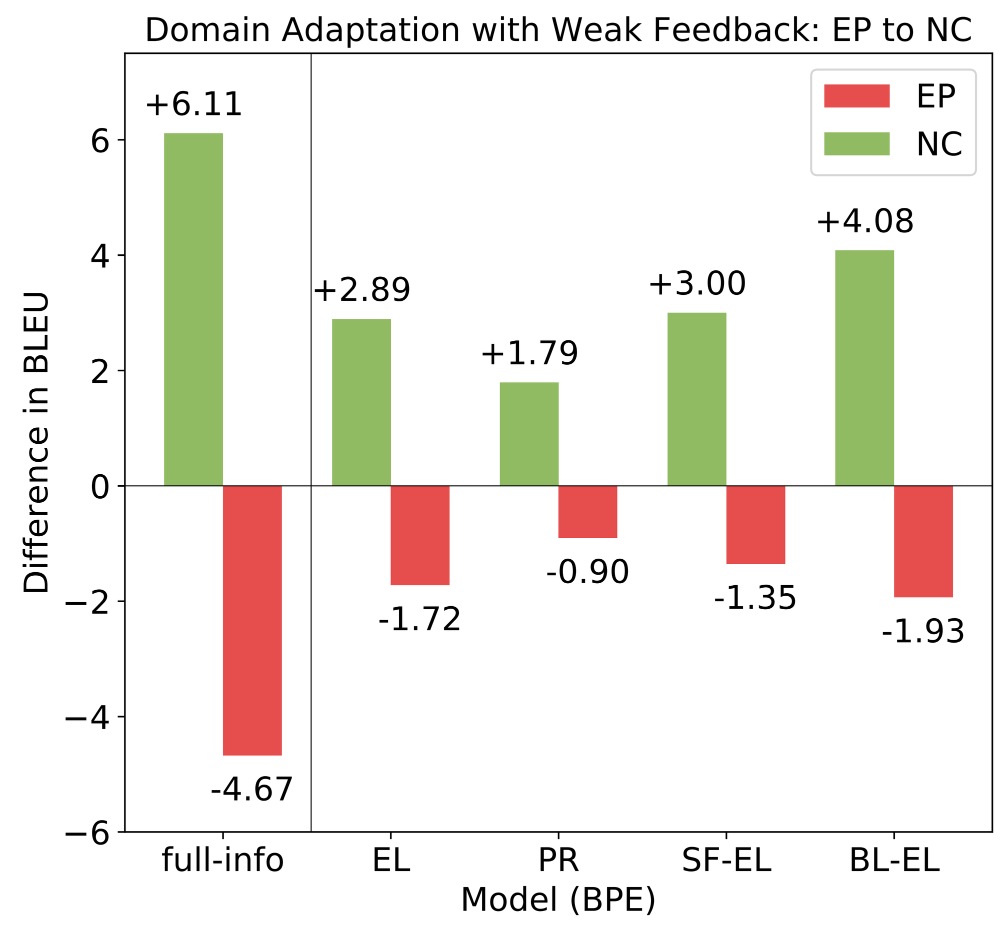
1 attention-based replacement of UNKs for word-based models [4]

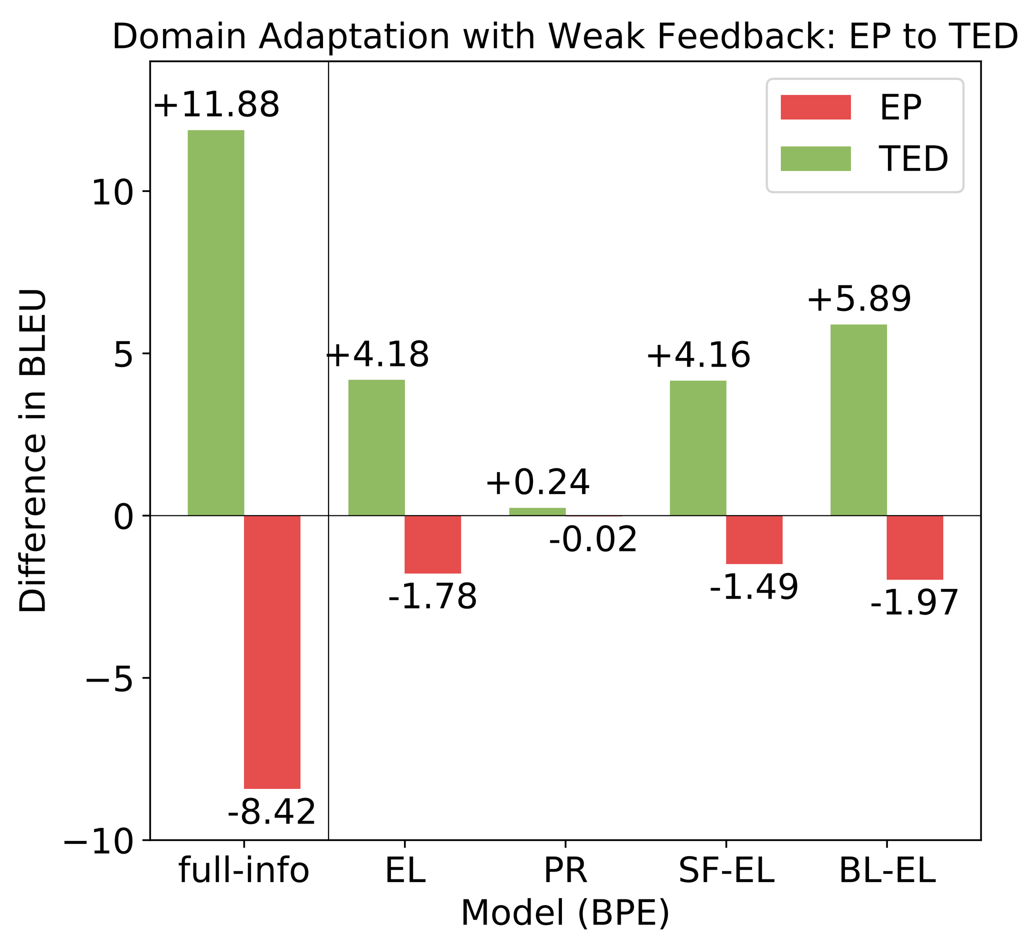
2 sub-word models with Byte-Pair-Encoding (BPE) [5]

Results

BLEU on held-out in- and out-of-domain test sets for parameters $\hat{\theta}$ selected by **early stopping** on a validation set.

We seek models for **conservative domain adaptation**, that learn to improve on in-domain, but maintain quality on out-of-domain translations.





Findings

- Successful training of NMT with weak feedback
- Large improvements for domain adaptation, outperforming linear models
- Control variates improve generalization, see [6]

Acknowledgements

This research was supported in part by the German research foundation (DFG), and in part by a research cooperation grant with the Amazon Development Center Germany.

References

[1] A. Sokolov, J. Kreutzer, C. Lo, and S. Riezler. Stochastic structured prediction under bandit feedback. In *NIPS*, Barcelona, Spain, 2016.

[2] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 20:229–256, 1992.

[3] R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *AISTATS*, Reykjavik, Iceland, 2014.

[4] S. Jean, O. Firat, K. Cho, R. Memisevic, and Y. Bengio. Montreal neural machine translation systems for WMT'15. In *WMT*, Lisbon, Portugal, 2015.

[5] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. In *ACL*, Berlin, Germany, 2016.

[6] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, New York, NY, 2016.