Bandit Structured Prediction for Learning from Partial Feedback in SMT

Artem Sokolov[†], Stefan Riezler[‡], Tanguy Urvoy[*]

[†‡]Computational Linguistics and [‡]IWR, Heidelberg University, Germany

[*]Orange Labs, Lannion, France

Human post-editing is limited

- by cost of professional translators
- by the required user expertise
- by time constraints, e.g., simultaneous translation

Human post-editing is limited

- by cost of professional translators
- by the required user expertise
- by time constraints, e.g., simultaneous translation

Adaptive Learning of SMT from post-edits is limited by

- unclear mapping of post-edits to SMT operations
- small number of post-edits

Human post-editing is limited

- by cost of professional translators
- by the required user expertise
- by time constraints, e.g., simultaneous translation

Adaptive Learning of SMT from post-edits is limited by

- unclear mapping of post-edits to SMT operations
- small number of post-edits

Goals

- adaptive learning of SMT from partial feedback
- faster and easier interaction (for user and learner)
- long-term goal: personalized adaptive SMT!

A look at online advertising

- estimate click-through-rate (CTR) for ads
- tradeoff between **exploration** (display new ad) and **exploitation** (display ad with current best estimate of CTR)
- only **one-point/bandit feedback** (click on ad/pull arm of slot machine) available for learning

A look at online advertising

- estimate click-through-rate (CTR) for ads
- tradeoff between **exploration** (display new ad) and **exploitation** (display ad with current best estimate of CTR)
- only **one-point/bandit feedback** (click on ad/pull arm of slot machine) available for learning

Online learning from bandit feedback

1. observe input structure x_t
2. sample output structure y_t
3. receive feedback to sampled structure, e.g., task loss at point y_t
4. update parameters

Partial feedback: learner does not know correct structure nor what would have happened if it had predicted differently!

Contributions

- Theory
  - ➡ algorithm for structured prediction from bandit feedback
  - ➡ applied to expected loss objective (Och, 2003; Smith and Eisner, 2006; Gimpel and Smith, 2010)
  - ➡ convergence analysis in the stochastic optimization framework of pseudogradient adaptation (Polyak and Tsypkin, 1973)
- Practice
  - ➡ simulated bandit feedback: evaluate task loss (BLEU) against reference only for sampled structure
  - ➡ re-ranking experiment: improve out-of-domain SMT model based on bandit feedback from in-domain data by 1.26 to 1.52 BLEU points

- $\mathcal{X}$ – structured input space
- $\mathcal{Y}(x)$ – set of possible output structures for x
- $\Delta_y(y') : \mathcal{Y} \to [0, 1]$ – loss suffered for predicting y' instead of y
- underlying Gibbs distribution

$$p_w(y|x) = \frac{\exp(w^\top \phi(x, y))}{Z_w(x)},$$

- $\mathcal{X}$ – structured input space
- $\mathcal{Y}(x)$ – set of possible output structures for x
- $\Delta_y(y') : \mathcal{Y} \to [0, 1]$ – loss suffered for predicting y' instead of y
- underlying Gibbs distribution

$$p_w(y|x) = \frac{\exp(w^\top \phi(x, y))}{Z_w(x)},$$

EL objective

$$\mathbb{E}_{p(x,y)p_w(y'|x)} \left[\Delta_y(y')\right] = \sum_{x,y} p(x, y) \sum_{y' \in \mathcal{Y}(x)} \Delta_y(y') p_w(y'|x)$$

- $\mathcal{X}$ – structured input space
- $\mathcal{Y}(x)$ – set of possible output structures for x
- $\Delta_y(y') : \mathcal{Y} \rightarrow [0,1]$ – loss suffered for predicting y' instead of y
- underlying Gibbs distribution

$$p_w(y|x) = \frac{\exp(w^\top \phi(x,y))}{Z_w(x)},$$

EL objective

$$\mathbb{E}_{p(x,y)p_w(y'|x)}\left[\Delta_y(y')\right] = \sum_{x,y} p(x,y) \sum_{y' \in \mathcal{Y}(x)} \Delta_y(y') p_w(y'|x)$$

Inference

- MBR: $\hat{y}_w(x) = \arg\min_{y \in \mathcal{Y}(x)} \sum_{y' \in \mathcal{Y}(x)} \Delta_y(y') p_w(y'|x)$
- MAP: $\hat{y}_w(x) = \arg\max_{y \in \mathcal{Y}(x)} p_w(y|x)$

Full information case

- $p(x, y)$ can be approximated by the empirical distribution $\tilde{p}(x, y)$

$$\mathbb{E}_{\tilde{p}(x,y)p_w(y'|x)} \left[\Delta_y(y')\right] = \frac{1}{T} \sum_{t=0}^{T} \sum_{y' \in \mathcal{Y}(x_t)} \Delta_{y_t}(y') p_w(y'|x_t)$$

- continuous and differentiable, but typically non-convex
- still, most approaches rely on gradient-descent techniques

Full information case

- $p(x, y)$ can be approximated by the empirical distribution $\tilde{p}(x, y)$

$$\mathbb{E}_{\tilde{p}(x,y)p_w(y'|x)}\left[\Delta_y(y')\right] = \frac{1}{T}\sum_{t=0}^{T}\sum_{y'\in\mathcal{Y}(x_t)} \Delta_{y_t}(y')p_w(y'|x_t)$$

- continuous and differentiable, but typically non-convex
- still, most approaches rely on gradient-descent techniques

Gradient

$$\begin{aligned}
&\nabla\mathbb{E}_{\tilde{p}(x,y)p_w(y'|x)}\left[\Delta_y(y')\right] \\
&\quad = \mathbb{E}_{\tilde{p}(x,y)}\Big[\mathbb{E}_{p_w(y'|x)}[\Delta_y(y')\phi(x, y')] - \mathbb{E}_{p_w(y'|x)}[\Delta_y(y')]\ \mathbb{E}_{p_w(y'|x)}[\phi(x, y')]\Big] \\
&\quad = \mathbb{E}_{\tilde{p}(x,y)p_w(y'|x)}\Big[\Delta_y(y')(\phi(x, y') - \mathbb{E}_{p_w(y'|x)}[\phi(x, y')])\Big]
\end{aligned}$$

Bandit feedback means that the gold standard y is not revealed

- we can neither calculate the gradient of the objective function
- nor evaluate the task loss Δ as in the full information case

Bandit feedback means that the gold standard y is not revealed

- we can neither calculate the gradient of the objective function
- nor evaluate the task loss Δ as in the full information case

- solution
  - ➡ pass the evaluation of $\Delta(y)$ to the user
  - ➡ slightly change the objective

$$J(w) = \mathbb{E}_{p(x)p_w(y'|x)}\left[\Delta(y')\right] = \sum_x p(x) \sum_{y' \in \mathcal{Y}(x)} \Delta(y') p_w(y'|x)$$

  - ➡ reflects that we don't model unseen y
  - ➡ natural for SMT: no single true translation

Algorithm 1 Bandit Structured Prediction

1: Input: sequence of learning rates γ_t
2: Initialize w_0
3: **for** $t = 0, \ldots, T$ **do**
4: Observe x_t
5: Calculate $\mathbb{E}_{p_{w_t}(y'|x_t)}[\phi(x_t, y')]$
6: Sample $\tilde{y}_t \sim p_{w_t}(y'|x_t)$
7: Obtain feedback $\Delta(\tilde{y}_t)$
8: Update $w_{t+1} = w_t - \gamma_t \, \Delta(\tilde{y}_t)\Big(\phi(x_t, \tilde{y}_t) - \mathbb{E}_{p_{w_t}(y'|x_t)}[\phi(x_t, y')]\Big)$

- simultaneous exploration/exploitation by sampling from Gibbs distribution
- compare the feature vector of $\tilde{y}_t$ to the average feature vector
- step into opposite direction of this difference, depending on feedback $\Delta(\tilde{y}_t)$
- step size is bigger for high loss
- extreme case: no update if $\tilde{y}_t$ is correct, i.e. $\Delta(\tilde{y}_t) = 0$

Pseudogradient adaptation framework (Polyak and Tsypkin, 1973)

- iterative process/algorithm

$$w_{t+1} = w_t - \gamma_t \, s_t \tag{1}$$

  - ➡ $\gamma_t \geq 0$ is a learning rate
  - ➡ w_t and s_t are random vectors in $\mathbb{R}^d$
  - ➡ the distribution of s_t depends on $w_0, \ldots, w_t$

Pseudogradient adaptation framework (Polyak and Tsypkin, 1973)

- iterative process/algorithm

$$w_{t+1} = w_t - \gamma_t \, s_t \tag{1}$$

  - ➡ $\gamma_t \geq 0$ is a learning rate
  - ➡ w_t and s_t are random vectors in $\mathbb{R}^d$
  - ➡ the distribution of s_t depends on $w_0, \ldots, w_t$

Pseudogradient condition

Random vector s_t is a *pseudogradient* of an objective $J(w)$ if

$$\nabla J(w_t)^\top \mathbb{E}[s_t] \geq 0,$$

i.e., s_t is on average at an acute angle with $\nabla J(w)$.

Technical conditions:

- boundedness of the update vector

$$\mathbb{E}[||s_t||^2] < \infty$$

- learning rate does not decrease too fast

$$\gamma_t \geq 0, \ \sum_{t=0}^{\infty} \gamma_t = \infty, \ \sum_{t=0}^{\infty} \gamma_t^2 < \infty$$

- $J(w)$ is lower bounded and differentiable
- gradient $\nabla J(w)$ is Lipschitz continuous s.t. for all w, w', there exists $L \geq 0$, such that

$$\left|\left|\nabla J(w + w') - \nabla J(w)\right|\right| \leq L \left|\left|w'\right|\right|$$

Theorem (Polyak and Tsypkin (1973), Thm. 1)

Under the above conditions, for any starting w_0 *in process* (1)*:*

$$J(w_t) \to J^* \;\; almost\ surely,\ and\ \lim_{t\to\infty} \nabla J(w_t)^\top \mathbb{E}(s_t) = 0.$$

Theorem (Polyak and Tsypkin (1973), Thm. 1)

Under the above conditions, for any starting w_0 in process (1):

$$J(w_t) \to J^* \;\; \textit{almost surely, and} \; \lim_{t \to \infty} \nabla J(w_t)^\top \mathbb{E}(s_t) = 0.$$

Significance

- conditions can be checked easily
- no need to know the gradient on every step to verify the condition
- applies to a wide range of cases, including non-convex functions (convergence to a critical point)

Algorithm 1 as stochastic approximation algorithm

$$J(w) = \mathbb{E}_{p(x)p_w(y'|x)}[\Delta(y')]$$

$$\nabla J(w) = \mathbb{E}_{\tilde{p}(x)p_w(y'|x)}\left[\Delta_y(y')(\phi(x,y') - \mathbb{E}_{p_w(y'|x)}[\phi(x,y')])\right]$$

$$s_t = \Delta(\tilde{y}_t)(\phi(x_t,\tilde{y}_t) - \mathbb{E}_{p_{w_t}(y|x_t)}[\phi(x_t,y)])$$

- pseudogradient condition holds since s_t is an unbiased estimate of the true gradient s.t. $\mathbb{E}_{p(x)p_{w_t}(y'|x)}[s_t] = \nabla J(w_t)$ and

$$\nabla J(w_t)^\top \mathbb{E}_{p(x)p_{w_t}(y'|x)}[s_t] = ||\nabla J(w_t)||^2 \geq 0$$

- assuming $||\phi(x,y')|| \leq R$ and $\Delta(y') \in [0,1]$ for all x, y':

$$\mathbb{E}_{p(x)p_{w_t}(y'|x)}[||s_t||^2] \leq 4R^2$$

- decreasing learning rate, e.g. $\gamma_t = 1/t$

Algorithm 2 Structured Dueling Bandits

Input: γ, δ, w_0
for $t = 0, \ldots, T$ **do**
    Observe x_t
    Sample unit vector u_t uniformly
    Set $w'_t = w_t + \delta u_t$
    Compare $\Delta(\hat{y}_{w_t}(x_t))$ to $\Delta(\hat{y}_{w'_t}(x_t))$
    if w'_t wins **then**
        $w_{t+1} = w_t + \gamma u_t$
    else
        $w_{t+1} = w_t$

- generic algorithm (Yue and Joachims, 2009) applied to structured prediction
- explicit control of exploration (δ) and exploitation (γ)
- requires much stronger two-point feedback

General setup

- simulated bandit feedback by evaluating task loss against gold-standard structures *without* revealing them to the learner
- online learning for parameter estimation
- online-to-batch conversion of last model at test time
- results on the test set under MAP inference
- final results averaged over 5 independent runs

SMT reranking setup

- **idea:** Simulate personalized SMT by adapting out-of-domain system to a user by single-point bandit feedback
- **simulation:** SMT domain adaptation by 5k-best list reranking using simulated bandit feedback from in-domain data

Data

- WMT'07 shared task, Europarl to NewsCommentary, FR-EN
- out-of-domain parallel data: 1.6 million Europarl
- in-domain parallel data: train/dev/test: 43,194/1,064/2,007 NewsCommentary
- language model for both: Europarl target side + **in-domain** NewsCommentary

Models

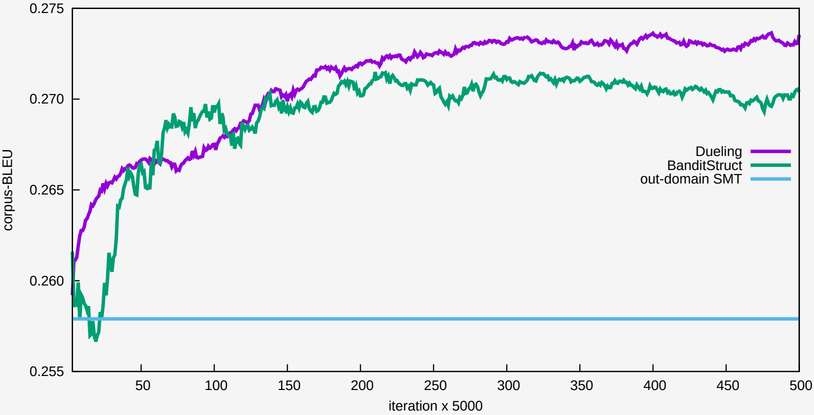
- phrase-based (?), 4-gram language model (?), 15 dense features

Tuning

- **full information:**
  - ➡ MERT tuning on out-of-domain or in-domain dev set, respectively
  - ➡ MERT runs repeated 7 times, median result reported
- **bandit learning:**
  - ➡ online bandit learning on in-domain train set, started from out-of-domain median model, smoothed per-sentence 1-BLEU task loss
  - ➡ in-domain dev set for meta-parameter tuning (learning rate, exploration parameter)
  - ➡ testing by online-to-batch conversion of last model after 100 epochs by corpus-BLEU on in-domain test set

full information		bandit information	
in-domain SMT	**out-domain SMT**	**DuelingBandit**	**BanditStruct**
0.2854	0.2579	$0.2731_{\pm 0.001}$	$0.2705_{\pm 0.001}$

- BanditStruct and DuelingBandit very close, despite the latter is using twice as much information
- both are considerable improvements over out-of-domain model (remember: out-domain SMT uses in-domain lm!)
  - ➡ BanditStruct: $+1.26$ BLEU points
  - ➡ DuelingBandit: $+1.52$ BLEU points
- all results statistically significant



- per-sentence BLEU is a difficult metric for bandit feedback
- smoother and faster convergence curve for Dueling Bandits since relative information can be exploited

Conclusion

- convergent algorithm for structured prediction from single-point feedback
- promising empirical results, both compared to two-point feedback and to full information scenarios
- strength where correct structures are unavailable and two-point feedback is infeasible

Current and future work

- other "banditizable" objectives for structured prediction
  - ➡ pairwise preference learning under single-point feedback
  - ➡ strongly convex objective for improved convergence rate
- real-world feedback
  - ➡ deployment in CAT course for translation students

Current and future work

- other “banditizable” objectives for structured prediction
  - ➡ pairwise preference learning under single-point feedback
  - ➡ strongly convex objective for improved convergence rate
- real-world feedback
  - ➡ deployment in CAT course for translation students

Thank you!

Gimpel, K. and Smith, N. A. (2010). Softmax-margin training for structured log-linear models. Technical Report CMU-LTI-10-008, Carnegie Mellon University, Pittsburgh, PA.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *HLT-NAACL*, Edmonton, Canada.

Polyak, B. T. and Tsypkin, Y. Z. (1973). Pseudogradient adaptation and training algorithms. *Automation and remote control*, 34(3):377–397.

Smith, D. A. and Eisner, J. (2006). Minimum risk annealing for training log-linear models. In *COLING-ACL*, Sydney, Australia.

Yue, Y. and Joachims, T. (2009). Interactively optimizing information retrieval systems as a dueling bandits problem. In *ICML*, Montreal, Canada.