Bandit Structured Prediction for Learning from Partial Feedback in Statistical Machine Translation

Artem Sokolov sokolov@cl.uni-heidelberg.de
Stefan Riezler* riezler@cl.uni-heidelberg.de
Computational Linguistics and IWR*
Heidelberg University, 69120 Heidelberg, Germany

Tanguy Urvoy tanguy.urvoy@orange.com
Orange Labs, 2 Avenue Pierre Marzin, 22307 Lannion, France

Abstract

We present an approach to structured prediction from bandit feedback, called *Bandit Structured Prediction*, where only the value of a task loss function at a single predicted point, instead of a correct structure, is observed in learning. We present an application to discriminative reranking in Statistical Machine Translation (SMT) where the learning algorithm only has access to a $1 -$ BLEU loss evaluation of a predicted translation instead of obtaining a gold standard reference translation. In our experiment bandit feedback is obtained by evaluating BLEU on reference translations without revealing them to the algorithm. This can be thought of as a simulation of interactive machine translation where an SMT system is personalized by a user who provides single point feedback to predicted translations. Our experiments show that our approach improves translation quality and is comparable to approaches that employ more informative feedback in learning.

1 Introduction

Learning from *bandit*[1] feedback describes an online learning scenario, where on each of a sequence of rounds, a learning algorithm makes a prediction, and receives partial information in terms of feedback to a single predicted point. In difference to the full information supervised scenario, the learner does not know what the correct prediction looks like, nor what would have happened if it had predicted differently. This scenario has (financially) important real world applications such as online advertising (Chapelle et al., 2014) that showcases a tradeoff between exploration (a new ad needs to be displayed in order to learn its click-through rate) and exploitation (displaying the ad with the current best estimate is better in the short term). Crucially, in this scenario it is unrealistic to expect more detailed feedback than a user click on the displayed ad. Similar to the online advertising scenario, there are many potential applications of bandit learning to NLP situations where feedback is limited for various reasons. For example, online learning has been applied successfully in interactive statistical machine translation (SMT) (Bertoldi et al., 2014; Denkowski et al., 2014; Green et al., 2014). Post-editing feedback clearly is limited by its high cost and by the required expertise of users, however, current approaches force the full information supervised scenario onto the problem of learning from user post-edits.

[1] The name is inherited from a model where in each round a gambler pulls an arm of a different slot machine ("one-armed bandit"), with the goal of maximizing his reward relative to the maximal possible reward, without apriori knowledge of the optimal slot machine.

Bandit learning would allow to learn from partial user feedback that is easier and faster to obtain than full information. An example where user feedback is limited by a time constraint is simultaneous translation of a speech input stream (Cho et al., 2013). Clearly, it is unrealistic to expect user feedback that goes beyond a one-shot user quality estimate of the predicted translation in this scenario. Another example is SMT domain adaptation where the translations of a large out-of-domain model are re-ranked based on bandit feedback on in-domain data. This can also be seen as a simulation of personalized machine translation where a given large SMT system is adapted to a user solely by single-point user feedback to predicted structures.

The goal of this paper is to develop algorithms for structured prediction from bandit feedback, tailored to NLP problems. We investigate possibilities to "banditize" objectives such as expected loss (Och, 2003; Smith and Eisner, 2006; Gimpel and Smith, 2010) that have been proposed for structured prediction in NLP. Since most current approaches to bandit optimization rely on a multiclass classification scenario, the first challenge of our work is to adapt bandit learning to structured prediction over exponentially large structured output spaces (Taskar et al., 2004; Tsochantaridis et al., 2005). Furthermore, most theoretical work on online learning with bandit feedback relies on convexity assumptions about objective functions, both in the non-stochastic adversarial setting (Flaxman et al., 2005; Shalev-Shwartz, 2012) as well as in the stochastic optimization framework (Spall, 2003; Nemirovski et al., 2009; Bach and Moulines, 2011). Our case is a non-convex optimization problem, which we analyze in the simple and elegant framework of pseudogradient adaptation that allows us to show convergence of the presented algorithm (Polyak and Tsypkin, 1973; Polyak, 1987).

The central contributions of this paper are:

- An algorithm for minimization of expected loss for structured prediction from bandit feedback, called *Bandit Structured Prediction*.
- An analysis of convergence of our algorithm in the stochastic optimization framework of pseudogradient adaptation.
- An experimental evaluation on structured learning in SMT. Our experiment follows a simulation design that is standard in bandit learning, namely by simulating bandit feedback by evaluating task loss functions against gold standard structures without revealing them to the learning algorithm.

As a disclaimer, we would like to note that improvements over traditional full-information structured prediction cannot be expected from learning from partial feedback. Instead, the goal is to investigate learning situations in which full information is not available. Similarly, a comparison between our approach and dueling bandits (Yue and Joachims, 2009) is skewed towards the latter approach that has access to two-point feedback instead of one-point feedback as in our case. While it has been shown that querying the loss function at two points leads to convergence results that closely resemble bounds for the full information case (Agarwal et al., 2010), such feedback is clearly twice as expensive and, depending on the application, might not be elicitable from users.

2 Related Work

Stochastic Approximation. Online learning from bandit feedback dates back to Robbins (1952) who formulated the task as a problem of sequential decision making. His analysis, as ours, is in a stochastic setting, i.e., certain assumptions are made on the probability distribution of feedback and its noisy realization. Stochastic approximation covers bandit feedback as noisy observations which only allow to compute noisy gradients that equal true gradients in expectation. While the stochastic approximation framework is quite general, most theoretical analyses of convergence and convergence rate are based on (strong) convexity assumptions (Polyak and Juditsky, 1992; Spall, 2003; Nemirovski et al., 2009; Bach and Moulines, 2011, 2013) and thus not applicable to our case.

Non-Stochastic Bandits. Auer et al. (2002) initiated an active area of research on non-stochastic bandit learning, i.e., no statistical assumptions are made on the distribution of feedback, including models of feedback as a malicious choice of an adaptive adversary. The adversarial bandit setting has been extended to take context or side information into account, using models based on general linear classifiers (Auer et al., 2002; Langford and Zhang, 2007; Chu et al., 2011). However, they formalize a multi-class classification problem that is not easily scalable to general exponentially large structured output spaces. Furthermore, most theoretical analyses rely on online (strongly) convex optimization (Flaxman et al., 2005; Shalev-Shwartz, 2012) thus limiting the applicability to our case.

Neurodynamic Programming. Bertsekas and Tsitsiklis (1996) cover optimization for neural networks and reinforcement learning under the name of "neurodynamic programming". Both areas are dealing with non-convex objectives that lead to stochastic iterative algorithms. Interestingly, the available analyses of non-convex optimization for neural networks and reinforcement learning in Bertsekas and Tsitsiklis (1996), Sutton et al. (2000), or Bottou (2004) rely heavily on Polyak and Tsypkin (1973)'s pseudogradient framework. We apply their simple and elegant framework directly to give asymptotic guarantees for our algorithm.

NLP Applications. In the area of NLP, recently algorithms for response-based learning have been proposed to alleviate the supervision problem by extracting supervision signals from task-based feedback to system predictions. For example, Goldwasser and Roth (2013) presented an online structured learning algorithm that uses positive executability of a semantic parse against a database to convert a predicted parse into a gold standard structure for learning. Riezler et al. (2014) apply a similar idea to SMT by using the executability of a semantic parse of a translated database query as signal to convert a predicted translation into gold standard reference in structured learning. Sokolov et al. (2015) present a coactive learning approach to structured learning in SMT where instead of a gold standard reference a slight improvement over the prediction is shown to be sufficient for learning. Saluja and Zhang (2014) present an incorporation of binary feedback into an latent structured SVM for discriminative SMT training. NLP applications based on reinforcement learning have been presented by Branavan et al. (2009) or Chang et al. (2015). Their model differs from ours in that it is structured as a sequence of states at which actions and rewards are computed, however, the theoretical foundation of both types of models can be traced back to Polyak and Tsypkin (1973)'s pseudogradient framework .

3 Expected Loss Minimization under Full Information

The expected loss learning criterion for structured prediction is defined as a minimization of the expectation of a task loss function with respect to the conditional distribution over structured outputs (Gimpel and Smith, 2010; Yuille and He, 2012). More formally, let $\mathcal{X}$ be a structured input space, let $\mathcal{Y}(x)$ be the set of possible output structures for input x, and let $\Delta_y : \mathcal{Y} \to [0, 1]$ quantify the loss $\Delta_y(y')$ suffered for making errors in predicting y' instead of y; as a rule, $\Delta_y(y') = 0$ iff $y = y'$. Then, for a data distribution $p(x, y)$, the learning criterion is defined as minimization of the expected loss

$$\mathbb{E}_{p(x,y)p_w(y'|x)}\left[\Delta_y(y')\right] = \sum_{x,y} p(x,y) \sum_{y' \in \mathcal{Y}(x)} \Delta_y(y') p_w(y'|x). \tag{1}$$

Assume further that output structures given inputs are distributed according to an underlying Gibbs distribution (a.k.a. conditional exponential or log-linear model)

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

where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a joint feature representation of inputs and outputs, $w \in \mathbb{R}^d$ is a weight vector, and $Z_w(w)$ is a normalization constant.

The natural rule for prediction or inference is according to the minimum Bayes risk principle

$$\hat{y}_w(x) = \arg\min_{y \in \mathcal{Y}(x)} \sum_{y' \in \mathcal{Y}(x)} \Delta_y(y') p_w(y'|x). \tag{2}$$

This requires an evaluation of $\Delta_y(y')$ over the full output space, which is standardly avoided in practice by performing inference according to a maximum a posteriori (MAP) criterion (which equals criterion (2) for the special case of $\Delta_y(y') = \mathbf{1}[y \neq y']$ where $\mathbf{1}[s]$ evaluates to 1 if statement s is true, 0 otherwise)

$$\begin{aligned} \hat{y}_w(x) &= \arg\max_{y \in \mathcal{Y}(x)} p_w(y|x) \\ &= \arg\max_{y \in \mathcal{Y}(x)} w^\top \phi(x,y). \end{aligned} \tag{3}$$

Furthermore, since it is unfeasible to take expectations over the full space $\mathcal{X} \times \mathcal{Y}$ to perform minimization of objective (1), in the full information case the data distribution $p(x,y)$ is approximated by the empirical distribution $\tilde{p}(x,y) = \frac{1}{T}\sum_{t=0}^{T} \mathbf{1}[x = x_t]\mathbf{1}[y = y_t]$ for i.i.d. training data $\{(x_t, y_t)\}_{t=0}^{T}$. This yields the objective

$$\mathbb{E}_{\tilde{p}(x,y)p_w(y'|x)}\left[\Delta_y(y')\right] = \frac{1}{T}\sum_{t=0}^{T} \sum_{y' \in \mathcal{Y}(x_t)} \Delta_{y_t}(y') p_w(y'|x_t). \tag{4}$$

While being continuous and differentiable, the expected loss criterion is typically non-convex. For example, in SMT, expected loss training for the standard task loss BLEU leads to highly non-convex optimization problems. Despite of this, most approaches rely on gradient-descent techniques for optimization (see Och (2003), Smith and Eisner (2006), He and Deng (2012), Auli et al. (2014), Wuebker et al. (2015), *inter alia*) by following the opposite direction of the gradient of (4):

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

4 Bandit Structured Prediction

Bandit feedback in structured prediction means that the gold standard output structure y, with respect to which the objective function is evaluated, is not revealed to the learner. Thus we can neither calculate the gradient of the objective function (4) nor evaluate the task loss Δ as in the full information case. A solution to this problem is to pass the evaluation of the loss function to the user, i.e, we access the loss directly through user feedback without assuming existence of a fixed reference y. We indicate this by dropping the subscript y in $\Delta(y')$. Assuming a fixed,

Algorithm 1 Bandit Structured Prediction

```
1: Input: sequence of learning rates $\gamma_t$
2: Initialize $w_0$
3: for $t = 0, \dots, T$ 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 $w_{t+1} = w_t - \gamma_t \ \Delta(\tilde{y}_t)(\phi(x_t, \tilde{y}_t) - \mathbb{E}_{p_{w_t}(y'|x_t)}[\phi(x_t, y')])$
```

unknown distribution $p(x)$ over input structures, we can formalize the following new objective for expected loss minimization in a bandit setup

$$\begin{aligned} J(w) &= \mathbb{E}_{p(x)p_w(y'|x)} \left[\Delta(y')\right] \\ &= \sum_x p(x) \sum_{y' \in \mathcal{Y}(x)} \Delta(y') p_w(y'|x). \end{aligned} \tag{5}$$

Optimization of this objective is then as follows:

1. We assume a sequence of input structures $x_t, t = 1, \dots, T$ that are generated by a fixed, unknown distribution $p(x)$.

2. We use a Gibbs distribution estimate as a sampling distribution to perform simultaneous exploration / exploitation on output structures (Abernethy and Rakhlin, 2009).

3. We use feedback to the sampled output structures to construct a parameter update rule that is an unbiased estimate of the true gradient of objective (5).

4.1 Algorithm

Algorithm 1 implements these ideas as follows: We assume as input a given learning rate schedule (line 1) and a deterministic initialization w_0 of the weight vector (line 2). For each random i.i.d. input structure x_t, we calculate the expected feature count (line 5). This can be done exactly, provided the underlying graphical model permits a tractable calculation, or for intractable models, with MCMC sampling. We then sample an output structure $\tilde{y}_t$ from the Gibbs model (line 6). If the number of output options is small, this is done by sampling from a multinomial distribution. Otherwise, we use a Perturb-and-MAP approach (Papandreou and Yuille, 2011), restricted to unary potentials, to obtain an approximate Gibbs sample without waiting for the MC chain to mix. Finally, an update in the negative direction of the instantaneous gradient, evaluated at the input structure x_t (line 8), is performed.

Intuitively, the algorithm compares the sampled feature vector to the average feature vector, and performs a step into the opposite direction of this difference, the more so the higher the loss of the sampled structure is. In the extreme case, if the sampled structure is correct ($\Delta(\tilde{y}_t) = 0$), no update is performed.

4.2 Stochastic Approximation Analysis

The construction of the update in Algorithm 1 as a stochastic realization of the true gradient allows us to analyze the algorithm as a stochastic approximation algorithm. We show how our case can be fit in the pseudogradient adaptation framework of Polyak and Tsypkin (1973) which gives asymptotic guarantees for non-convex and convex objectives. They characterize an iterative process

$$w_{t+1} = w_t - \gamma_t \; s_t \tag{6}$$

where $\gamma_t \geq 0$ is a learning rate, w_t and s_t are vectors in $\mathbb{R}^d$ with fixed w_0, and the distribution of s_t depends on $w_0, \ldots, w_t$. For a given lower bounded and differentiable function $J(w)$ with Lipschitz continuous gradient $\nabla J(w)$, that is, for all w, w', there exists $L \geq 0$, such that

$$\|\nabla J(w + w') - \nabla J(w)\| \leq L\|w'\|, \tag{7}$$

the vector s_t in process (6) is said to be a *pseudogradient* of $J(w)$ if

$$\nabla J(w_t)^\top \mathbb{E}[s_t] \geq 0, \tag{8}$$

where the expectation is taken over all sources of randomness. Intuitively, the pseudogradient s_t is on average at an acute angle with the true gradient, meaning that $-s_t$ is on average a direction of decrease of the functional $J(w)$.

In order to show convergence of the iterative process (6), besides conditions (7) and (8), only mild conditions on boundedness of the pseudogradient

$$\mathbb{E}[\|s_t\|^2] < \infty, \tag{9}$$

and on the use of a decreasing learning rate satisfying

$$\gamma_t \geq 0, \; \sum_{t=0}^{\infty} \gamma_t = \infty, \; \sum_{t=0}^{\infty} \gamma_t^2 < \infty, \tag{10}$$

are necessary. Under the exclusion of trivial solutions such as $s_t = \mathbf{0}$, the following convergence assertion can be made:

Theorem 1 (Polyak and Tsypkin (1973), Thm. 1) *Under conditions* (7)–(10)*, for any* w_0 *in process* (6)*:*

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

The significance of the theorem is that its conditions can be checked easily, and it applies to a wide range of cases, including non-convex functions, in which case the convergence point J^* is a critical point of $J(w)$.

The convergence analysis of Theorem 1 can be applied to Algorithm 1 as follows: First note that we can define our functional $J(w)$ with respect to expectations over the full space of $\mathcal{X}$ as $J(w) = \mathbb{E}_{p(x)p_w(y'|x)}[\Delta(y')]$. This means, convergence of the algorithm can be understood directly as a generalization result that extends to unseen data. In order to show this result, we have to verify conditions (7)–(10). It is easy to show that condition (7) holds for our functional $J(w)$. Next we match the update in Algorithm 1 to a vector

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

Taking the expectation of s_t yields $\mathbb{E}_{p(x)p_{w_t}(y'|x)}[s_t] = \nabla J(w_t)$ such that condition (8) is satisfied by

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

Assuming $\|\phi(x, y')\| \leq R$ and $\Delta(y') \in [0, 1]$ for all x, y', condition (9) is satisfied by

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

For a decreasing learning rate, e.g., $\gamma_t = 1/t$, condition (10) holds, such that convergence to a critical point of the expected risk follows according to Theorem 1.

Algorithm Structured Dueling Bandits

```
1:  Input: $\gamma, \delta, w_0$
2:  for $t = 0, \dots, T$ do
3:      Observe $x_t$
4:      Sample unit vector $u_t$ uniformly
5:      Set $w'_t = w_t + \delta u_t$
6:      Compare $\Delta(\hat{y}_{w_t}(x_t))$ to $\Delta(\hat{y}_{w'_t}(x_t))$
7:      if $w'_t$ wins then
8:          $w_{t+1} = w_t + \gamma u_t$
9:      else
10:         $w_{t+1} = w_t$
```

5 Structured Dueling Bandits

For purposes of comparison, we present an extension of Yue and Joachims (2009)'s dueling bandits algorithm to structured prediction problems. The original algorithm is not specifically designed for structured prediction problems, but it is generic enough to be applicable to such problems when the quality of a parameter vector can be proxied through loss evaluation of an inferred structure.

The Structured Dueling Bandits algorithm compares a current weight vector w_t with a neighboring point w'_t along a direction u_t, performing exploration (controlled by δ, line 5) by probing random directions, and exploitation (controlled by γ, line 8) by taking a step into the winning direction. The comparison step in line 6 is adapted to structured prediction from the original algorithm of Yue and Joachims (2009) by comparing the quality of w_t and w'_t via an evaluation of the losses $\Delta(\hat{y}_{w_t}(x_t))$ and $\Delta(\hat{y}_{w'_t}(x_t))$ of the structured arms corresponding to MAP prediction (3) under w_t and w'_t, respectively.

Further, note that the Structured Dueling Bandit algorithm requires access to a two-point feedback instead of a one-point feedback as in case of Bandit Structured Prediction (Algorithm 1). It has been shown that two-point feedback leads to convergence results that are close to those for learning from full information Agarwal et al. (2010). However, two-point feedback is twice as expensive as one-point feedback, and most importantly, such feedback might not be elicitable from users in real-world situations where feedback is limited by time- and resource-constraints. This limits the range of applications of Dueling Bandits to real-world interactive scenarios.

6 Experiments

Our experimental design follows the standard of simulating bandit feedback by evaluating task loss functions against gold standard structures without revealing them to the learner. We compare the proposed Structured Bandit Prediction algorithm to Structured Dueling Bandits, and report results by test set evaluations of the respective loss functions under MAP inference. Furthermore, we evaluate models at different iterations according to their loss on the test set in order to visualize the empirical convergence behavior of the algorithms.

All experiments with bandit algorithms perform online learning for parameter estimation, and apply early stopping to choose the last model in a learning sequence for online-to-batch conversion at test time. Final results for bandit algorithms are averaged over 5 independent runs.

In this experiment, we present bandit learning for the structured $1 - \text{BLEU}$ loss used in SMT. The setup is a reranking approach to SMT domain adaptation where the k-best list of an out-of-domain model is re-ranked (without re-decoding) based on bandit feedback from in-

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

Table 1: Corpus BLEU (under MAP decoding) on test set for SMT domain adaptation from Europarl to NewsCommentary by k-best reranking.

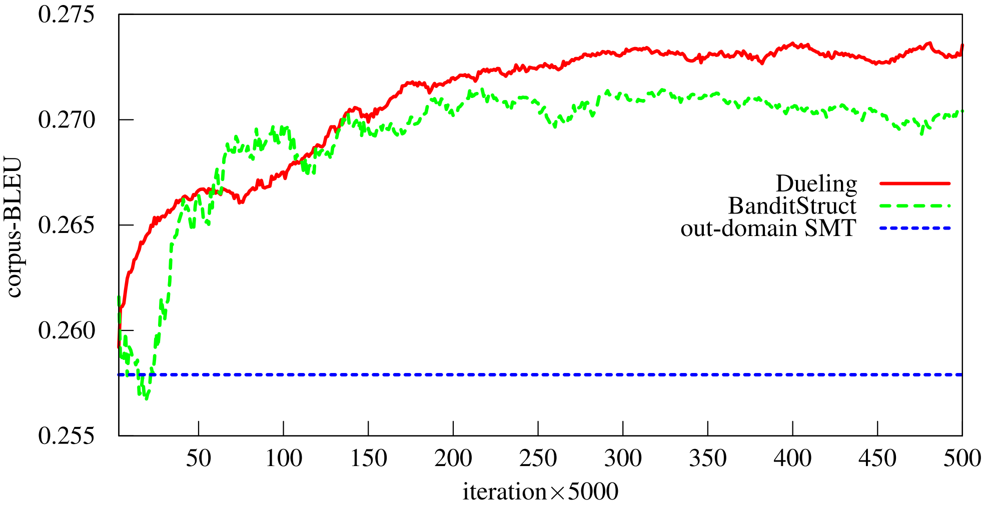

Figure 1: Corpus-BLEU on test set for early stopping at different iterations for the SMT task.

domain data. This can also be seen as a simulation of personalized machine translation where a given large SMT system is adapted to a user solely by single-point user feedback to predicted structures.

We use the data from the WMT 2007 shared task for domain adaptation experiments in a popular benchmark setup from Europarl to NewsCommentary for French-to-English (Koehn and Schroeder, 2007; Daumé and Jagarlamudi, 2011). We tokenized and lowercased our data using the `moses` toolkit, and prepared word alignments by `fast_align` (Dyer et al., 2013). The SMT setup is phrase-based translation using non-unique 5,000-best lists from `moses` (Koehn et al., 2007) and a 4-gram language model (Heafield et al., 2013).

The *out-of-domain* baseline SMT model is trained on 1.6 million parallel Europarl data and includes the English side of Europarl and *in-domain* NewsCommentary in the language model. The model uses 15 dense features (6 lexicalized reordering features, 1 distortion, 1 out-of-domain and 1 in-domain language model, 1 word penalty, 5 translation model features) that are tuned with MERT (Och, 2003) on a dev set of Europarl data (`dev2006`, 2,000 sentences). The full-information *in-domain* SMT model gives an upper bound by MERT tuning the out-of-domain model on in-domain development data (`nc-dev2007`, 1,057 sentences). MERT runs for both baseline models were repeated 7 times and median results are reported.

Learning under bandit feedback started at the learned weights of the *out-of-domain* median model. It uses the parallel NewsCommentary data (`news-commentary`, 43,194 sentences) to simulate bandit feedback, by evaluating the sampled translation against the gold standard reference using as loss function Δ a smoothed per-sentence $1 - \text{BLEU}$ (by flooring zero n-gram counts to 0.01). The meta-parameters of Dueling Bandits and Bandit Structured Prediction were adjusted by online optimization of cumulative per-sentence $1 - \text{BLEU}$ on a small *in-domain* dev set (`nc-devtest2007`, 1,064 parallel sentences). The final results are obtained by online-to-batch conversion where the model trained for 100 epochs on 43,194 *in-domain* training data is evaluated on a separate *in-domain* test set (`nc-test2007`, 2,007 sentences).

Table 1 shows that the results for Bandit Structured Prediction and Dueling Bandits are very close, however, both are significant improvements over the out-of-domain SMT model that even includes an in-domain language model. We show the standard evaluation of the corpus-BLEU metric evaluated under MAP inference. The range of possible improvements is given by the difference of the BLEU score of the in-domain model and the BLEU score of the out-of-domain model – nearly 3 BLEU points. Bandit learning can improve the out-of-domain baseline by about 1.26 BLEU points (Bandit Structured Prediction) and by about 1.52 BLEU points (Dueling Bandits). All result differences are statistically significant at a p-value of 0.0001, using an Approximate Randomization test (Riezler and Maxwell, 2005; Clark et al., 2011). Figure 1 shows that per-sentence BLEU is a difficult metric to provide single-point feedback, yielding a non-smooth progression of loss values against iterations for Bandit Structured Prediction. The progression of loss values is smoother and empirical convergence speed is faster for Dueling Bandits since it can exploit preference judgements instead of having to trust real-valued feedback.

7 Discussion

We presented an approach to *Bandit Structured Prediction* that is able to learn from feedback in form of an evaluation of a task loss function for *single* predicted structures. Our experimental evaluation showed promising results, both compared to Structured Dueling Bandits that employ two-point feedback, and compared to full information scenarios where the correct structure is revealed.

Our approach shows its strength where correct structures are unavailable and two-point feedback is infeasible. In future work we would like to apply bandit learning to scenarios with limited human feedback such as the interactive SMT applications discussed above. In such scenarios, per-sentence BLEU might not be the best metric to quantify feedback. We will instead investigate feedback based on HTER (Snover et al., 2006), or based on judgements according to Likert scales (Likert, 1932).

Acknowledgements

This research was supported in part by DFG grant RI-2221/2-1 “Grounding Statistical Machine Translation in Perception and Action”.

References

Abernethy, J. and Rakhlin, A. (2009). An efficient bandit algorithm for $\sqrt{T}$ regret in online multiclass prediction? In *Conference on Learning Theory (COLT)*, Montreal, Canada.

Agarwal, A., Dekel, O., and Xiao, L. (2010). Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Conference on Learning Theory (COLT)*, Haifa, Israel.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002). The nonstochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Auli, M., Galley, M., and Gao, J. (2014). Large-scale expected BLEU training of phrase-based reordering models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, Granada, Spain.

Bach, F. and Moulines, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, CA, USA.

Bertoldi, N., Simianer, P., Cettolo, M., Wäschle, K., Federico, M., and Riezler, S. (2014). Online adaptation to post-edits for phrase-based statistical machine translation. *Machine Translation*, 29:309–339.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific.

Bottou, L. (2004). Stochastic learning. In Bousquet, O. and von Luxburg, U., editors, *Advanced Lectures on Machine Learning*, pages 146–168.

Branavan, S., Chen, H., Zettlemoyer, L. S., and Barzilay, R. (2009). Reinforcement learning for mapping instructions to actions. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Suntec, Singapore.

Chang, K.-W., Krishnamurthy, A., Agarwal, A., Daume, H., and Langford, J. (2015). Learning to search better than your teacher. In *International Conference on Machine Learning (ICML)*, Lille, France.

Chapelle, O., Manavaglu, E., and Rosales, R. (2014). Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology*, 5(4).

Cho, E., Fügen, C., Hermann, T., Kilgour, K., Mediani, M., Mohr, C., Niehues, J., Rottman, K., Saam, C., Stüker, S., and Waibel, A. (2013). A real-world system for simultaneous translation of German lectures. In *Interspeech*, Lyon, France.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA.

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR.

Daumé, H. and Jagarlamudi, J. (2011). Domain adaptation for machine translation by mining unseen words. In *Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, Portland, OR, USA.

Denkowski, M., Dyer, C., and Lavie, A. (2014). Learning from post-editing: Online model adaptation for statistical machine translation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA, USA.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. (2005). Online convex optimization in the bandit setting: gradient descent without a gradient. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Philadelphia, PA.

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, USA.

Goldwasser, D. and Roth, D. (2013). Learning from natural instructions. *Machine Learning*, 94(2):205–232.

Green, S., Wang, S. I., Chuang, J., Heer, J., Schuster, S., and Manning, C. D. (2014). Human effort and machine learnability in computer aided translation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

He, X. and Deng, L. (2012). Maximum expected BLEU training of phrase and lexicon translation models. In *Meeting of the Association for Computational Linguistics (ACL)*, Jeju Island, Korea.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Koehn, P., Hoang, H., Birch, A., Callison-Birch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL Demo and Poster Sessions*, Prague, Czech Republic.

Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Langford, J. and Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*. Vancouver, Canada.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:5–55.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, Edmonton, Canada.

Papandreou, G. and Yuille, A. (2011). Perturb-and-map random fields: Using discrete optimization to learn and sample from energy models. In *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain.

Polyak, B. T. (1987). *Introduction to Optimization*. Optimization Software, Inc., New York.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

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

Riezler, S. and Maxwell, J. (2005). On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI.

Riezler, S., Simianer, P., and Haas, C. (2014). Response-based learning for grounded machine translation. In *Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, MD, USA.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Statistical Society*, 55:527–535.

Saluja, A. and Zhang, Y. (2014). Online discriminative learning for machine translation with binary-valued feedback. *Machine Translation*, 28:69–90.

Shalev-Shwartz, S. (2012). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

Smith, D. A. and Eisner, J. (2006). Minimum risk annealing for training log-linear models. In *International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, Sydney, Australia.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, MA, USA.

Sokolov, A., Riezler, S., and Cohen, S. B. (2015). A coactive learning view of online structured prediction in statistical machine translation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, Beijing, China.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley.

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.

Taskar, B., Klein, D., Collins, M., Koller, D., and Manning, C. (2004). Max-margin parsing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 5:1453–1484.

Wuebker, J., Muehr, S., Lehnen, P., Peitz, S., and Ney, H. (2015). A comparison of update strategies for large-scale maximum expected bleu training. In *Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies (NAACL-HLT)*, Denver, CO, USA.

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

Yuille, A. and He, X. (2012). Probabilistic models of vision and max-margin methods. *Frontiers of Electrical and Electronic Engineering*, 7(1).