Counterfactual Learning from Bandit Feedback under Deterministic Logging: A Case Study in SMT

Carolin Lawrence[1], Artem Sokolov[1,2], Stefan Riezler[1]

1 Department of Computational Linguistics, Heidelberg University, Germany. 2 Amazon Development Center, Germany.

Introduction

Commercial MT systems can easily log explicit or implicit feedback from users. Feedback is only available for one prediction, references are not available (**bandit setup**). The log is **biased** by the predictions of the logging system.

Counterfactual learning theory offers options to solve this problem but requires a **stochastic** logging system. Instead, commercial MT systems want to output only the most likely translation and are thus **deterministic**.

We show through clever usage of **control variates** that learning is possible despite of this. In domain-adaptation experiments with simulated feedback, we can report improvements of up to **2 BLEU**. Further, we can show that deterministic experiments are on a par with their stochastic counterparts.

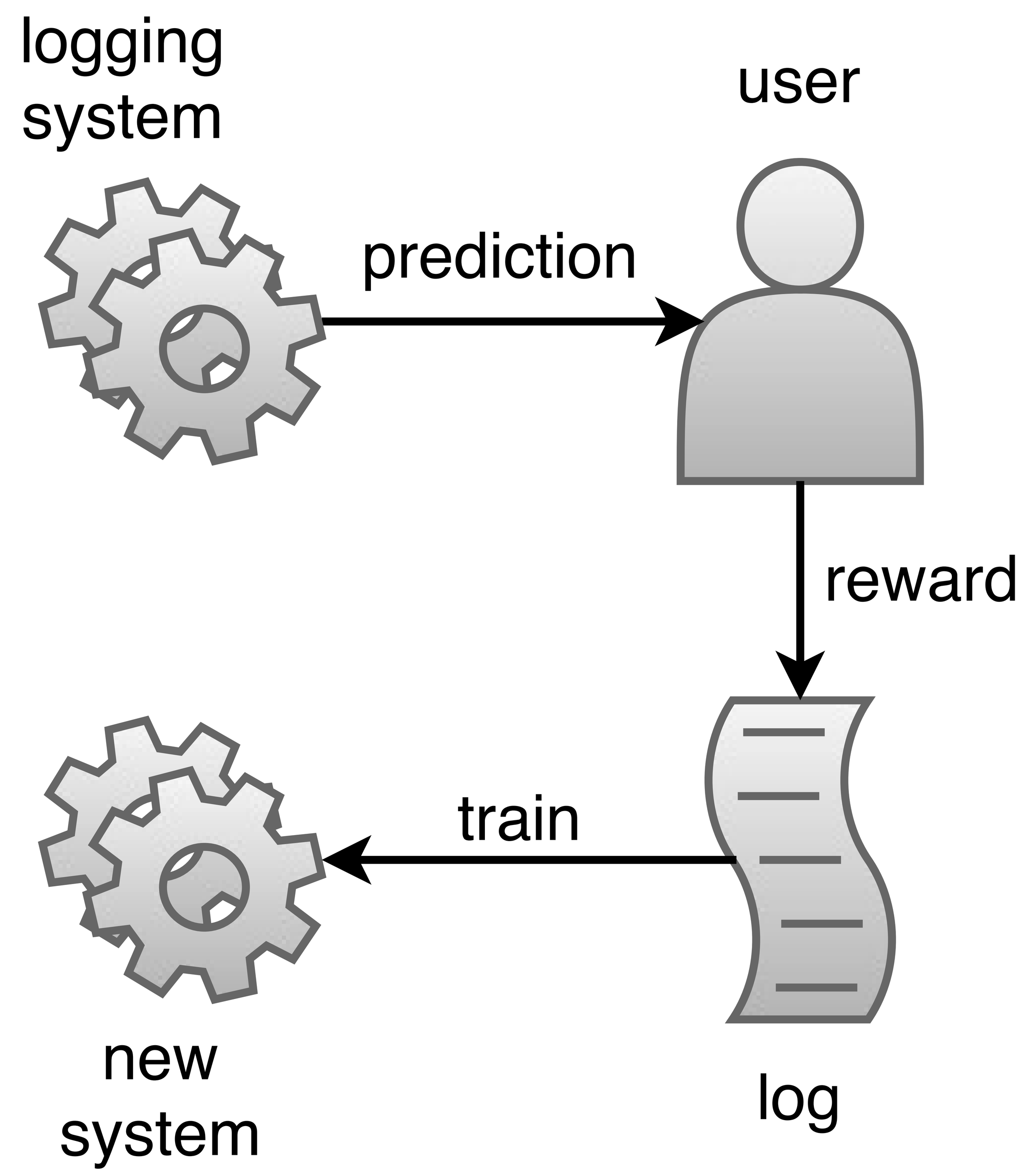


Figure 1: Offline learning from partial feedback.

Definitions

- collected: log $\mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$ where a logging system π_0 generated y_t given x_t and loss $\delta_t \in [-1, 0]$ is only given for the one y_t
- stochastic logging: record probability $\pi_0(y_t|x_t)$
- probability of current system: $\pi_w(y_t|x_t)$
- direct method (DM) predictor $\hat{\delta}$: can predict a reward for any input sequence

Objectives

Inverse Propensity Scoring (IPS)/ Deterministic Propensity Matching (DPM)

$\hat{R}_{\text{IPS/DPM}}(\pi_w) = \frac{1}{n}\sum_{t=1}^n \delta_t \rho_w(y_t|x_t)$

stochastic case

$\rho_w(y_t|x_t) = \frac{\pi_w(y_t|x_t)}{\pi_0(y_t|x_t)}$

- importance sampling corrects the bias in log
- y_t is sampled from the model distribution π_0 $\rightarrow$ exploration/exploitation trade-off

deterministic case

$\rho_w(y_t|x_t) = \pi_w(y_t|x_t)$ as $\pi_0(y_t|x_t) = 1$

Problem 1

- importance sampling is disabled
- y_t is the most likely translation under π_0 $\rightarrow$ exploration seems to be missing

Problem 2

- $\hat{R}_{\text{IPS/DPM}}(\pi_w)$ is maximized if all the log's probabilities $\pi_w(y_t|x_t)$ are set to 1 $\rightarrow$ increasing probability for low δ_t is undesired

Solution to 1

- **implicit exploration:** despite the deterministic logging, there is enough exploration because of the differing input context $\rightarrow$ deterministic logging can keep up with its stochastic counterpart

+ Multiplicative Control Variate:

Reweighting (+R)

Solution to 2

- define a probability distribution over the log $\rightarrow$ increasing probability for low δ_t will now decrease the objective as desired

$$\hat{R}_{\text{IPS+R/DPM+R}}(\pi_w) = \sum_{t=1}^n \delta_t \bar{\rho}_w(y_t|x_t) \quad (1)$$

with $\bar{\rho}_w(y_t|x_t) = \frac{\rho_w(y_t|x_t)}{\sum_t \rho_w(y_t|x_t)}$

Problem 3

- $\hat{R}_{\text{IPS+R/DPM+R}}(\pi_w)$ is maximized if the probability $\pi_w(y_t|x_t)$ of the highest δ_t is 1, the rest 0 $\rightarrow$ avoids logged data and potentially bad alternatives take up the probability mass of π_w

+ Additive Control Variate:

Doubly Robust (DR) / Doubly Controlled (DC)

Solution to 3

- use a DM predictor to evaluate the top scoring translations for each x_t $\rightarrow$avoiding logged data only possible if good alternatives take its place

$$\hat{R}_{\hat{c}\text{DR}/\hat{c}\text{DC}}(\pi_w) = \frac{1}{n}\sum_{t=1}^n \Big[(\delta_t - \hat{c}\hat{\delta}_t)\ \bar{\rho}_w(y_t|x_t) + \hat{c}\sum_{y\in\mathcal{Y}(x_t)} \hat{\delta}(x_t, y)\ \rho_w(y|x_t)\Big] \quad (3)$$

The optimal $\hat{c}$ can be derived: $\hat{c} = \frac{\text{Cov}(X,Y)}{\text{Var}(Y)}$

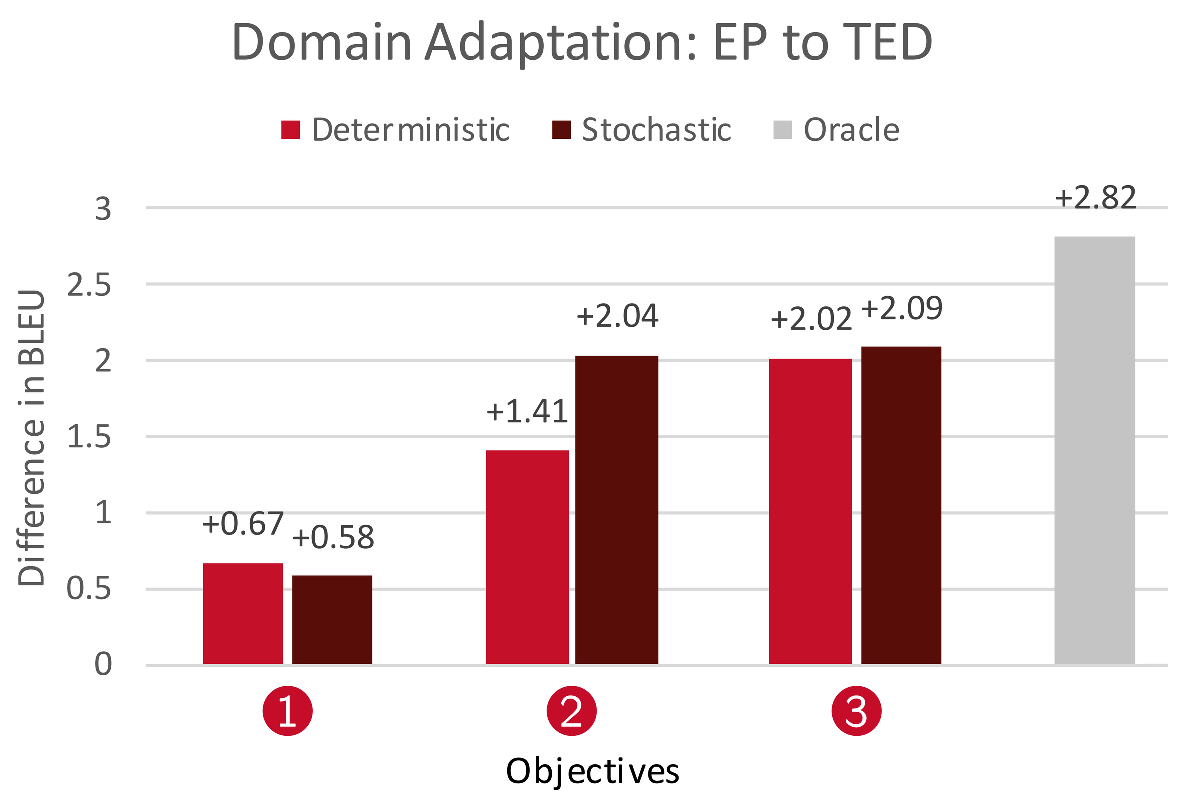
$\hat{R}_{\text{DR/DC}}(\pi_w)$ is $\hat{R}_{\hat{c}\text{DR}/\hat{c}\text{DC}}(\pi_w)$ with $\hat{c} = 1$ (2)
as defined by (Dudik et al., 2011)

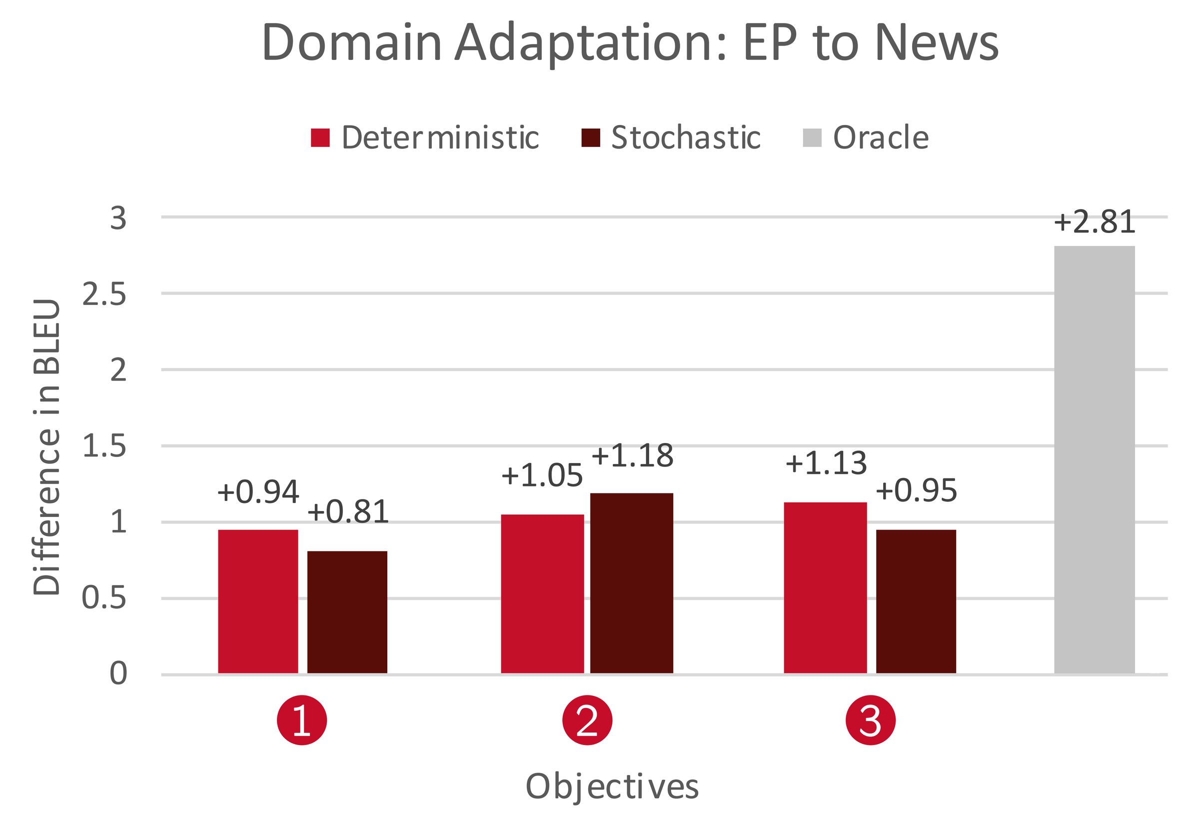
Experiments

Setup. Domain adaptation from Europarl (EP) to TED (de-en) & to News (fr-en) using phrase-based decoder CDEC & empirical risk minimization. Oracle systems where trained on references & MERT.

Log Creation. Logs were created by training a model on out-of-domain data & using this model to translate in-domain data. Feedback is simulated with negative per-sentence BLEU as the loss.

DM predictor $\hat{\delta}$. The predictor is a Scikit random forest model trained using the decoder's features as input & negative per-sentence BLEU as the output.





Take Away

- counterfactual learning works for MT despite large action space
- control variates fix problems of the simpler objectives
- $\hat{c}$ needs to be high to outperform setting $\hat{c} = 1$
- deterministic logging as good as stochastic $\rightarrow$ great advantage for e-commerce MT

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.

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386