Counterfactual Learning for Machine Translation: Degeneracies and Solutions

Carolin Lawrence[1], Pratik Gajane[2], Stefan Riezler[1]

1 Heidelberg University, Germany. 2 INRIA SequeL / Orange labs, France.

Neural Information Processing Systems

What IF Workshop

Overview

Commercial **Machine Translation** (MT) systems can easily log explicit or implicit feedback from users. To avoid the risk of showing inferior translations, commercial MT systems want to employ exploration-free policies which only output the most likely translation and are thus **deterministic**.

We show that the inverse and reweighted propensity scoring estimators can lead to possible **degeneracies** in both stochastic and deterministic setups. Using doubly robust methods, these degeneracies can avoided.

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 due to **implicit exploration**.

Definitions

- collected: log $\mathcal{D} = \{(x_t, y_t, \delta_t)\}_{t=1}^n$ where a logging system μ generated y_t given x_t and a reward $\delta_t \in [0,1]$ is observed
- stochastic logging: record probability $\mu(y_t|x_t)$
- 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)

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

stochastic case

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

deterministic case

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

Problem 1

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

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 [1]

Problem 2

Theorem 1 $max_\pi \hat{V}_{\text{IPS}}$ and $max_\pi \hat{V}_{\text{DPM}}$ if $\forall(y_t, x_t, \delta_t) \in \mathcal{D} : \pi(y_t|x_t) = 1 \wedge \delta_t > 0$.

$\hat{V}_{\text{IPS/DPM}}(\pi_w)$ is at maximum if all entries in the log with non-zero rewards receive probability 1

$\rightarrow$ increasing probability for low δ_t is undesired

Solution to 2

+ Multiplicative Control Variate [4]:

Reweighting (+R)

define a probability distribution over the log

$\rightarrow$ increasing probability for low δ_t will now decrease the objective as desired

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

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

Definition Let $\mathcal{D}^{max} = max_\delta \mathcal{D}$, then $\mathcal{D} = \mathcal{D}^{max} \cup \mathcal{D} \backslash \mathcal{D}^{max}$.

Theorem 2 $max_\pi \hat{V}_{\text{IPS+R}}$ and $max_\pi \hat{V}_{\text{DPM+R}}$ if $\exists(x_t, y_t, \delta_{max}) \in \mathcal{D}^{max} : \pi_t \in (0,1] \wedge \forall(y_t, x_t, \delta_t) \in \mathcal{D}\backslash\mathcal{D}^{max} : \pi_t = 0$.

$\hat{V}_{\text{IPS+R/DPM+R}}(\pi_w)$ is at maximum if the probability $\pi_w(y_t|x_t)$ of the highest δ_t is greater than 0 and the rest is 0

$\rightarrow$ avoids logged data and potentially bad alternatives take up the probability mass of π_w

Solution to 3

+ Additive Control Variate [2]:

Doubly Robust (DR) / Doubly Controlled (DC)

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

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

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

$\hat{V}_{\text{DR/DC}}(\pi_w)$ is $\hat{V}_{\hat{c}\text{DR}/\hat{c}\text{DC}}(\pi_w)$ with $\hat{c} = 1$ as defined by [2]. 2

Experiments [3]

Translation System. A Gibbs model that, given an input sentence x_t, defines probability distribution over all possible output sentences y_t,

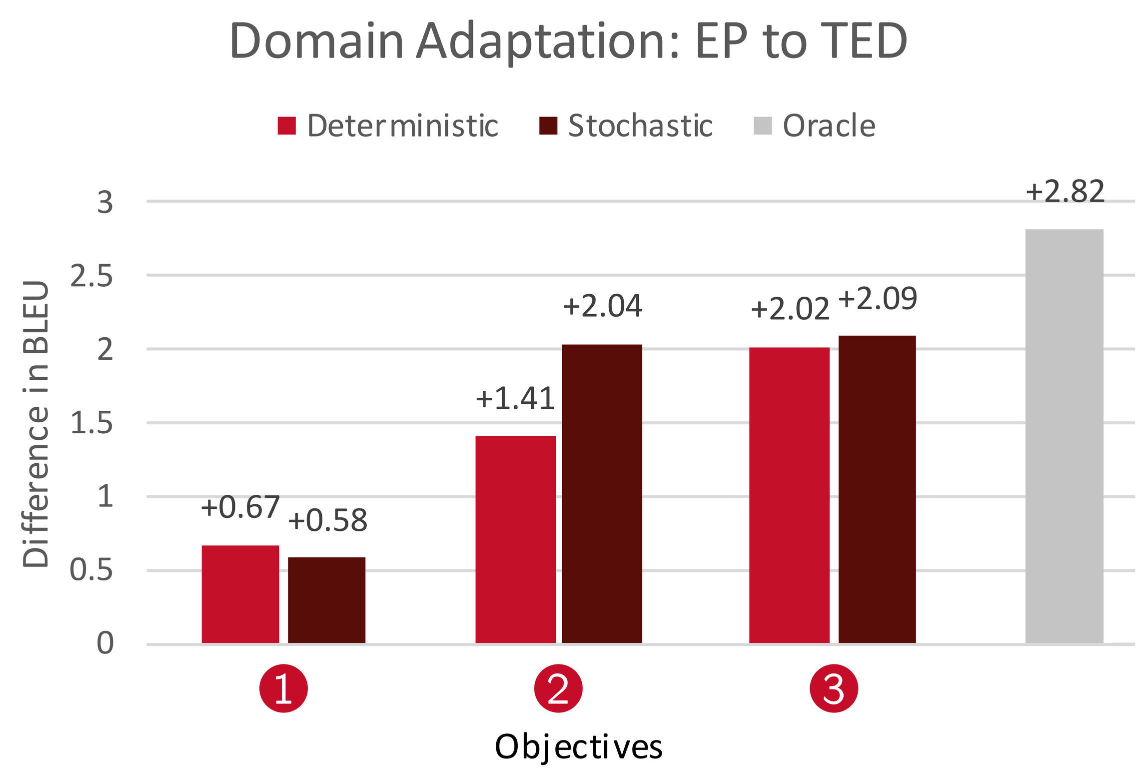
$$\pi_w(y_t|x_t) = \frac{e^{\alpha(w^\top \phi(x_t, y_t))}}{\sum_{y\in\mathcal{Y}(x_t)} e^{\alpha(w^\top \phi(x_t, y))}}.$$

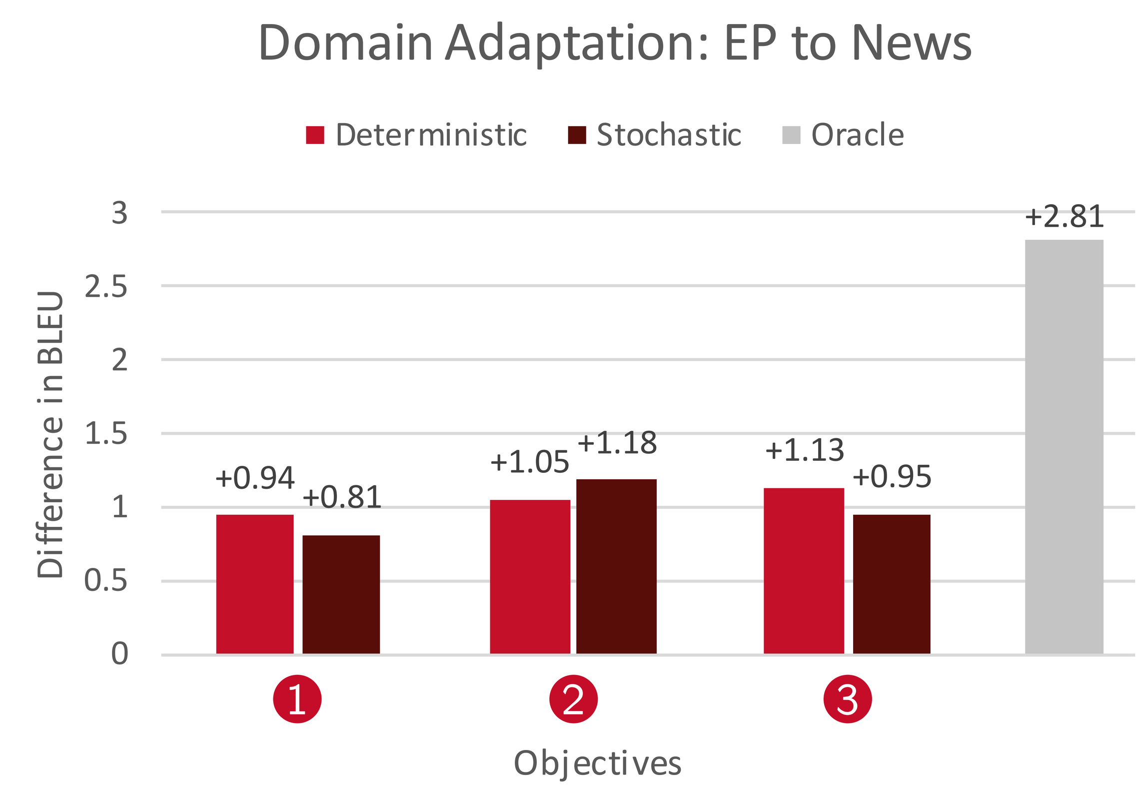
The number of possible output sentences may be very large. For example, assuming an output vocabulary of $90,000$ words and a sentence length of 200, there are $90,000^{200}$ possible outputs. Thus, the search for the most probable translation is often approximated, e.g. via beam search.

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

Log Creation. Logs were created by training a model on out-of-domain data and using this model to translate in-domain data. Feedback is simulated with per-sentence BLEU which is based on n-gram matching with regards to the gold translation.

DM predictor $\hat{\delta}$. The predictor is a Scikit random forest model trained using the decoder's features as input and 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
- deterministic logging as good as stochastic due to implicit exploration $\rightarrow$ great advantage for e-commerce MT

Acknowledgements

This research was supported in part by the German research foundation (DFG).

References

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

[2] Dudik, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *ICML*.

[3] Lawrence, C., Sokolov, A., and Riezler, S. (2017). Counterfactual learning from bandit feedback under deterministic logging: A case study in statistical machine translation. *EMNLP*.

[4] Swaminathan, A. and Joachims, T. (2015). The self-normalized estimator for counterfactual learning. In *NIPS*.