Neural Information Processing Systems

Counterfactual Learning from Human Proofreading Feedback for Semantic Parsing

Carolin Lawrence, Stefan Riezler

Heidelberg University, Germany.

LBI Workshop

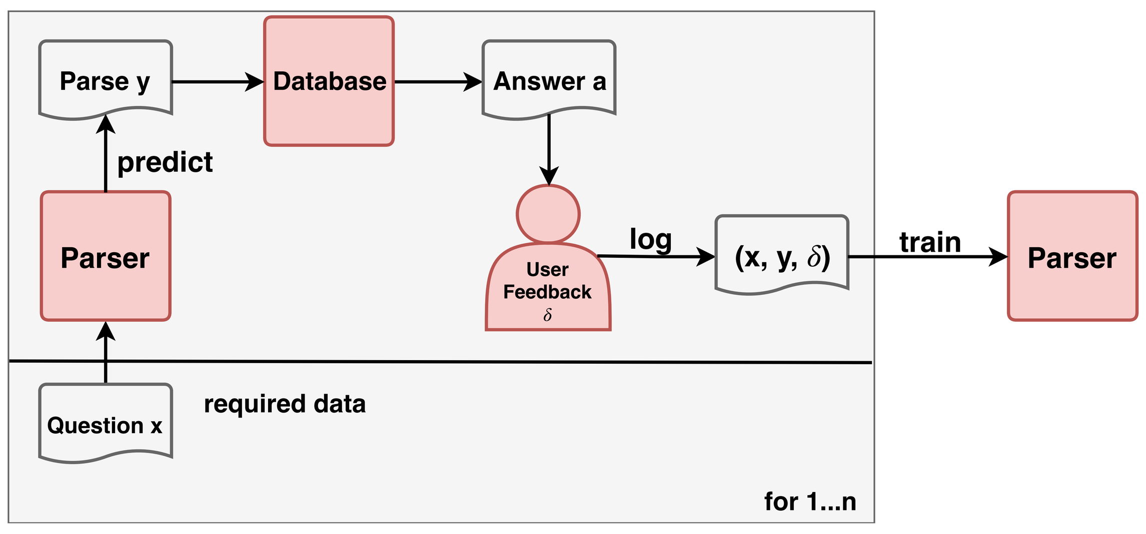
Overview

Training a semantic parser typically requires either **question-parse** pairs or **question-answer** pairs. Both can be expensive to obtain.

How can we alleviate the need for these pairs?

How can we further improve deployed parsers?

$\rightarrow$ **Collect feedback for model outputs from system-user interactions**



Difficult because

- No supervision: gold output is unknown
- Bandit: feedback for only one system output
- Bias: log $\mathcal{D}$ is biased to the decisions of the deployed parser

Solution: Counterfactual Off-policy Reinforcement Learning (CL)

Objectives

Collected log $\mathcal{D}_{log} = \{(x_t, y_t, \delta_t)\}_{t=1}^{n}$ with

- x_t: input
- y_t: most likely output of deployed system π_0
- $\delta_t \in [-1, 0]$: loss (i.e. neg. reward) from user

Deterministic Propensity Matching (DPM)

- Minimize expected risk for a target policy π_w

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

- Improve π_w using (stochastic) gradient descent

Problem: High variance

Solution: Multiplicative Control Variate - Reweighting (+R)

For random variables X and Y, with $\bar{Y}$ the expectation of Y:

$$\mathbb{E}[X] \approx \mathbb{E}[\frac{X}{Y}] \cdot \bar{Y}$$

$\rightarrow$ RHS has lower variance if Y positively correlates with X.

$$\hat{R}_{\text{DPM+R}}(\pi_w) = \frac{\frac{1}{n}\sum_{t=1}^{n} \delta_t \pi_w(y_t|x_t)}{\underbrace{\frac{1}{n}\sum_{t=1}^{n} \pi_w(y_t|x_t)}_{\text{Reweight Sum R}}} \cdot 1$$

- Reduces variance but introduces a bias of order $O(\frac{1}{n})$ that decreases as n increases
  $\rightarrow$ n should be as large as possible

Problem: To train state-of-the-art neural networks, stochastic minibatch learning is employed and then n is too small.

Task: NL Interface to OpenStreetMap (OSM)

- OSM: geographical database
- NLMAPS v2: 28,609 question-parse pairs
- Example question:

  "How many hotels are there in Paris?"

  Answer: 951
- Correctness of answers are difficult to judge
  $\rightarrow$ judge parses by making them human-understandable
- Feedback collection setup:
  1. Transform a parse to a set of statements
  2. Humans judge the statements

Automatically Transform a Parse

```
query(around(center(area(keyval('name','Paris')),
nwr(keyval('name','Place de la République'))),
search(nwr(keyval('amenity','parking'))),
maxdist(WALKING_DIST)),qtype(findkey('name')))
```

Question #216: **What are the names of cinemas that are within walking distance from the Place de la République in Paris?**

		Information found in Question?
Town	Paris	Yes No
Reference Point	name : Place de la République	Yes No
POI(s)	amenity : parking	Yes No
Question Type	What's the name	Yes No
Proximity	Around/Near	Yes No
Distance	Walking distance	Yes No

Submit

Note: If no question type is specified, the **default "Where"** is correct.

Solution:
One-Step Late (+OSL) Reweighting

Perform gradient descent updates & reweighting asynchronously:

- evaluate reweight sum R on the entire log of size n using past parameters w'
- update using minibatches of size m, $m \ll n$
- periodically update R

$\rightarrow$ retains all desirable properties

$$\hat{R}_{\text{DPM+OSL}}(\pi_w) = \frac{\frac{1}{m}\sum_{t=1}^{m} \delta_t \pi_w(y_t|x_t)}{\frac{1}{n}\sum_{t=1}^{n} \pi_{w'}(y_t|x_t)}$$

Problem:
Cannot learn from partially correct parses.

Solution: Token-Level (+T) Feedback

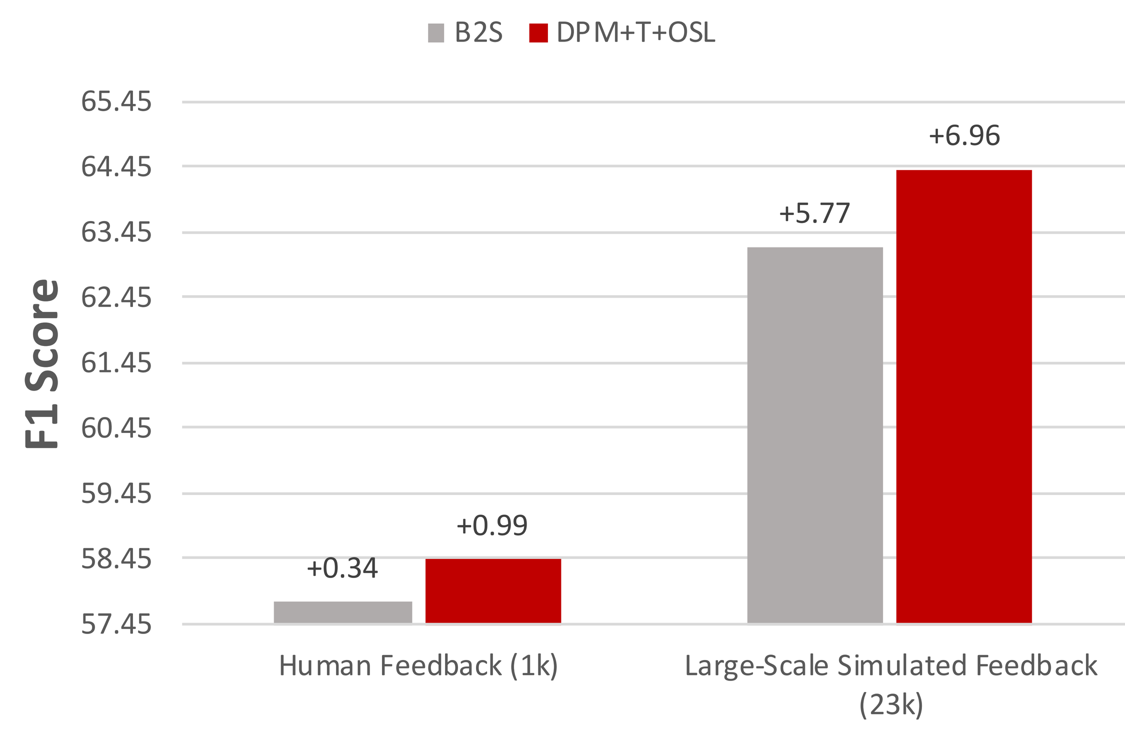
$$\hat{R}_{\text{DPM+T}}(\pi_w) = \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{j=1}^{|y|} \delta_j \log \pi_w(y_j|x_t)\right)$$

$$\hat{R}_{\text{DPM+T+OSL}}(\pi_w) = \frac{\frac{1}{m}\sum_{t=1}^{m}\left(\sum_{j=1}^{|y|} \delta_j \log \pi_w(y_j|x_t)\right)}{\frac{1}{n}\sum_{t=1}^{n} \pi_{w'}(y_t|x_t)}$$

Experiments

- Sequence-to-Sequence neural network NEMATUS
- Deployed system: pre-trained on 2k question-parse pairs
- Feedback collection:
  1. Humans judged 1k system outputs
     - Average time to judge a parse: 16.4s
     - Most parses (>70%) judged in <10s
  2. Simulated feedback for 23k system outputs
     - Token-wise comparison to gold parse
- Bandit-to-Supervised conversion (B2S): all instances in log with reward 1 are used as supervised training

B2S in comparison to the best CL objective:



Take Away

Proofreading

- Parses are automatically transformed into a set of human-understandable statements
- One set is typically judged in 10 seconds or less by a non-expert user
  $\rightarrow$ efficient alternative when the collection of question-parse or question-answer pairs is impossible or costly
- Feedback collection method enables blame assignment

Counterfactual Learning

- CL can safely improve models offline
- We introduce two new CL objectives:
  - DPM+OSL: a reweighting objective applicable to stochastic gradient optimization
  - DPM+T: effectively leverages the collected token-level feedback
- The combination DPM+T+OSL significantly outperforms a bandit-to-supervised baseline
- Can be applied to other tasks as well, e.g. machine translation

Future Work

Facilitate a dialogue with the user for a better user experience and to naturally encourage the collection of feedback.

Acknowledgements

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