Counterfactual Learning from Human Proofreading Feedback for Semantic Parsing

Carolin Lawrence, Stefan Riezler
Heidelberg University, Germany.

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?

→ Collect feedback from model outputs

from system-user interactions

Objectives

Collected log \( D_{\text{log}} = \{ (x_t, y_t, \delta_t) \}_{t=1}^{n} \)

- \( x_t \): input
- \( y_t \): most likely output of deployed system \( \pi_{0} \)
- \( \delta_t \in \{ -1, 0 \} \): loss (i.e., neg. reward) from user

Deterministic Propensity Matching (DPM)

- Minimize expected risk for a target policy \( \pi_{u} \)

\[
R_{\text{DPM}}(\pi_{u}) = \frac{1}{n} \sum_{t=1}^{n} \delta_t y_t(x_t) 
\]

- Improve \( \pi_{u} \) using (stochastic) gradient descent

Problem: High variance

Solution: Multiplicative Control Variate - Reweighting (+R)

For random variables \( X \) and \( Y \), with \( Y \) the expectation of \( Y \):

\[
E[X] \approx E \left[ \frac{X}{Y} \cdot Y \right] 
\]

→ RHS has lower variance if \( Y \) positively correlates with \( X \).

\[
\hat{R}_{\text{DPM}+R}(\pi_{u}) = \frac{1}{n} \sum_{t=1}^{n} \delta_t \pi_{u}(y_t|x_t) \cdot 1 
\]

Reweight Sum R

- Reduces variance but introduces a bias of order \( O(\frac{1}{n}) \) that decreases as \( n \) increases

→ \( n \) 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
- Solution: Counterfactual Off-policy Reinforcement Learning (CL)

Automatically Transform a Parse

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

Problem:

- CL can safely improve models offline
- We introduce two 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).

Take Away

- Parsers 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
- Efficient alternative when the collection of question-parse or question-answer pairs is impossible or costly
- Feedback collection method enables blame assignment

Counterfactual Learning

- Parse logs are transformed to collections of feedback.
- Counterfactual learning from human feedback (CL) can safely improve models offline.

Experiments

- Sequence-to-Sequence neural network Nematus
- Deployed system: pre-trained on 2k question-parse pairs
- Feedback collection:
  1. Humans judged 1k system outputs
  2. Average time to judge a parse: 16.4s
  3. 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:

Problem:

- 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
- Solution: Counterfactual Off-policy Reinforcement Learning (CL)

Automatically Transform a Parse

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

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 \( \theta' \)
- update using minibatches of size \( m \), \( m \ll n \)
- periodically update \( R \)

→ retains all desirable properties

\[
\hat{R}_{\text{DPM}+\text{OSL}}(\pi_{u}) = \frac{1}{n} \sum_{t=1}^{n} \delta_t \pi_{u}(y_t|x_t) 
\]

Problem:

- Cannot learn from partially correct parses.

Solution: Token-Level (+T) Feedback

\[
\hat{R}_{\text{DPM}+\text{T}}(\pi_{u}) = \frac{1}{n} \sum_{t=1}^{n} \delta_t \pi_{u}(y_t|x_t) 
\]

\[
\hat{R}_{\text{DPM}+\text{T}+\text{OSL}}(\pi_{u}) = \frac{1}{n} \sum_{t=1}^{n} \delta_t \pi_{u}(y_t|x_t) 
\]

- 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).