Simulating Bandit Learning from User Feedback for Extractive Question Answering

Ge Gao, Eunsol Choi, and Yoav Artzi

Overview

How to continually improve extractive QA systems?
- User feedback is an effective bandit learning signal
- Reduce data need and adapts to changing world
- Potential for domain adaption
- Simulation experiments with 6 existing supervised datasets

In-Domain Simulation

Scenario: very limited supervised training data
1) train an initial model on in-domain supervised data: 64 or 1024 examples
2) observe rewards and update the model on the fly
- Consistent performance gains on Wikipedia datasets
- Larger gains with weaker initial models
- Less effective with weaker initial models on datasets with noisy simulation

In-Domain: Online vs. Offline

Given the same initial model, compare online vs. offline setup:
- Offline learning is slightly more effective with stronger initial models on Wikipedia datasets
- Offline learning fails on noisier datasets even with stronger initial models

Domain Adaptation Simulation

Scenario: no supervised data available for the target domain
1) train an initial model on an existing dataset
2) adapt the model to new domain with bandit learning
- Performance gains on 22/30 configurations
- Extrapolates well particularly on HotpotQA from TriviaQA
- Effectiveness depends on the relation between domains

More in the paper:
- Sensitivity analysis to noisy user feedback
- Regret analysis: deficit suffered by the model relative to the optimal model
- Learning progression throughout the simulation

How to continually improve extractive QA systems?

In-Domain Simulation

Scenario: very limited supervised training data
1) train an initial model on in-domain supervised data: 64 or 1024 examples
2) observe rewards and update the model on the fly
- Consistent performance gains on Wikipedia datasets
- Larger gains with weaker initial models
- Less effective with weaker initial models on datasets with noisy simulation

In-Domain: Online vs. Offline

Given the same initial model, compare online vs. offline setup:
- Offline learning is slightly more effective with stronger initial models on Wikipedia datasets
- Offline learning fails on noisier datasets even with stronger initial models

Domain Adaptation Simulation

Scenario: no supervised data available for the target domain
1) train an initial model on an existing dataset
2) adapt the model to new domain with bandit learning
- Performance gains on 22/30 configurations
- Extrapolates well particularly on HotpotQA from TriviaQA
- Effectiveness depends on the relation between domains

More in the paper:
- Sensitivity analysis to noisy user feedback
- Regret analysis: deficit suffered by the model relative to the optimal model
- Learning progression throughout the simulation