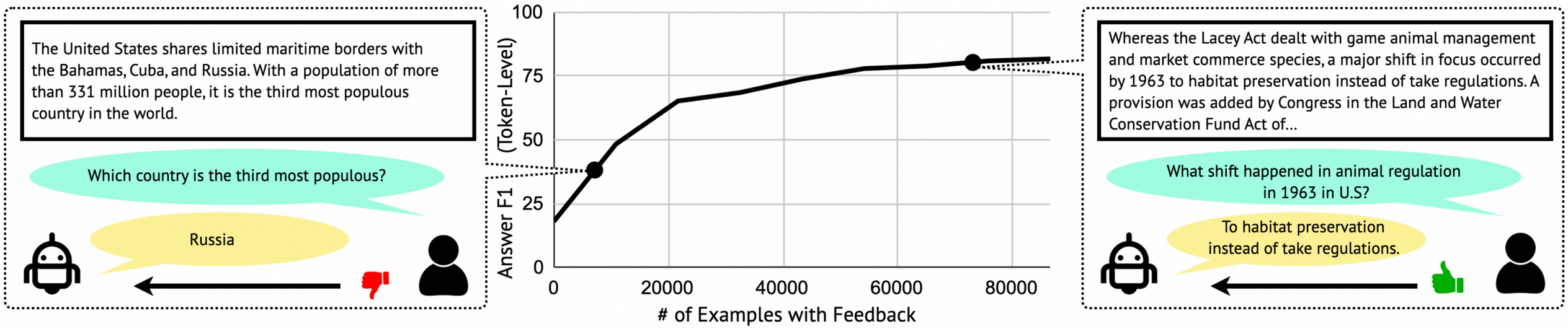
Simulating Bandit Learning from User Feedback for Extractive Question Answering

Ge Gao, Eunsol Choi, and Yoav Artzi

https://github.com/ lil-lab/bandit-qa

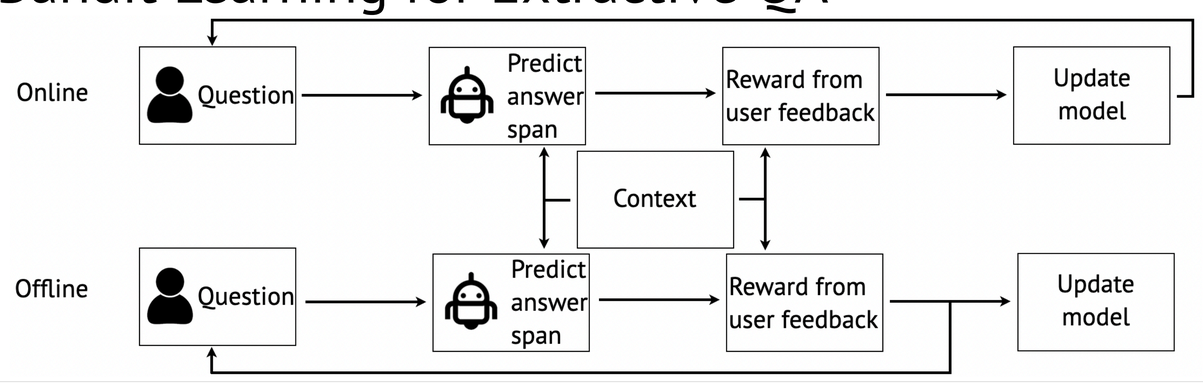


Overview

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

Bandit Learning for Extractive QA

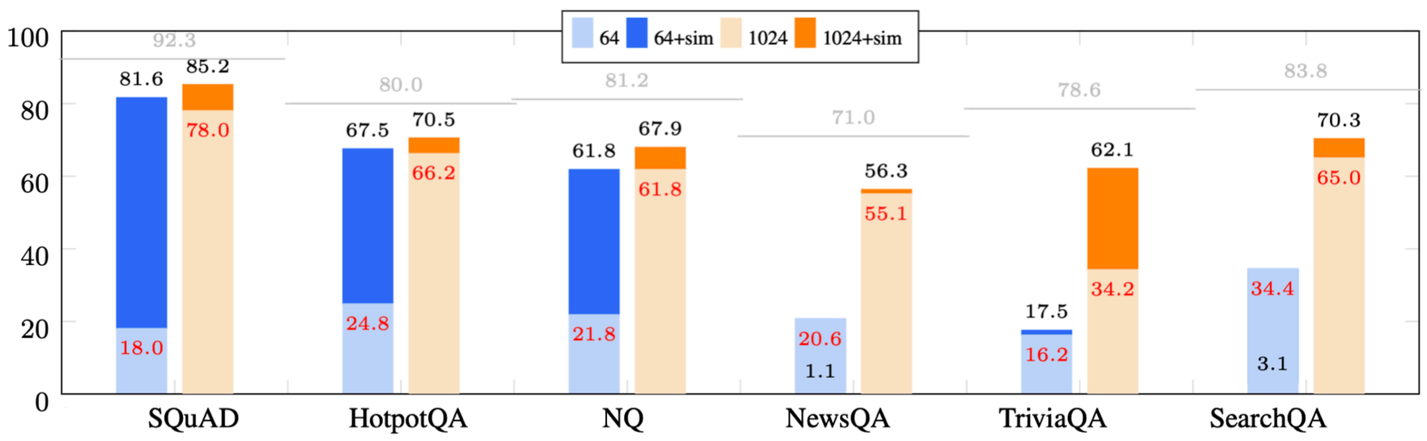


In-Domain Simulation

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



Red numbers: performance of initial models trained on 64 or 1024 examples. Black numbers: simulation performance. Horizontal grey lines: the supervised training performance.

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

Setup	SQuAD	HotpotQA	NQ	NewsQA	TriviaQA	SearchQA
64+sim	81.6 vs 78.2 [-3.4]	67.5 vs 66.3 [-1.2]	61.8 vs 51.3 [-10.5]	1.1 vs 3.1 [+2.0]	17.5 vs 0.4 [-17.1]	3.1 vs 1.3 [-1.8]
1024+sim	85.2 vs 86.5 [+1.3]	70.5 vs 73.2 [+2.7]	67.9 vs 71.8 [+3.9]	56.3 vs 55.7 [-0.6]	62.1 vs 7.5 [-54.6]	70.3 vs 4.1 [-66.2]

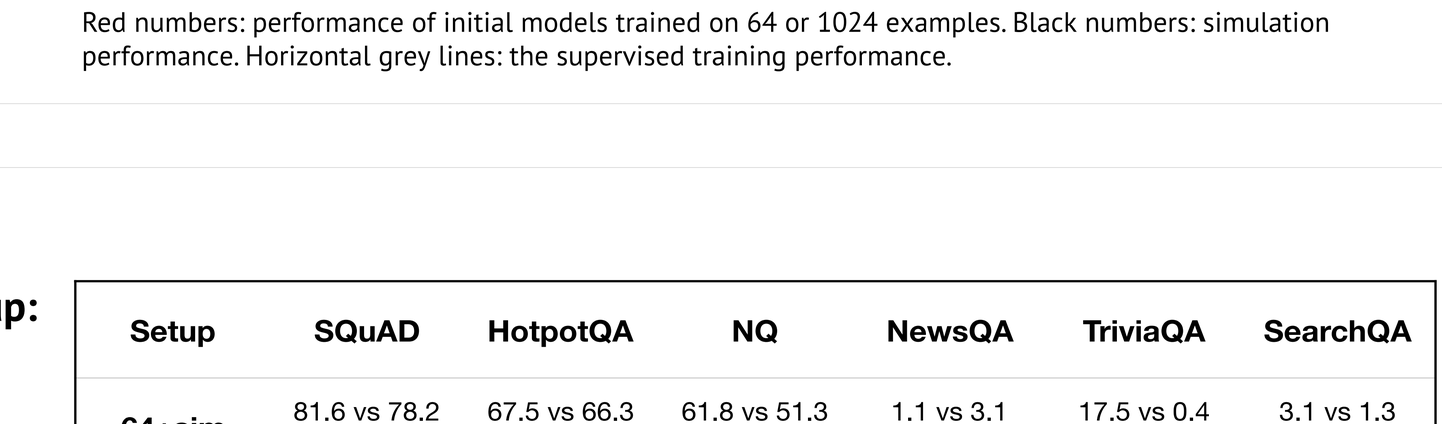
Online F1 vs offline F1. Colored numbers: offline – online.

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



Red numbers: performance of initial models trained on Y dataset (legend) on X dataset. Black numbers: simulation performance after adapting to X dataset.

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