Continually Improving Extractive QA via Human Feedback

Ge Gao*[1], Hung-Ting Chen*[2], Yoav Artzi[1], Eunsol Choi[2]

[1]Cornell University [2] University of Texas at Austin

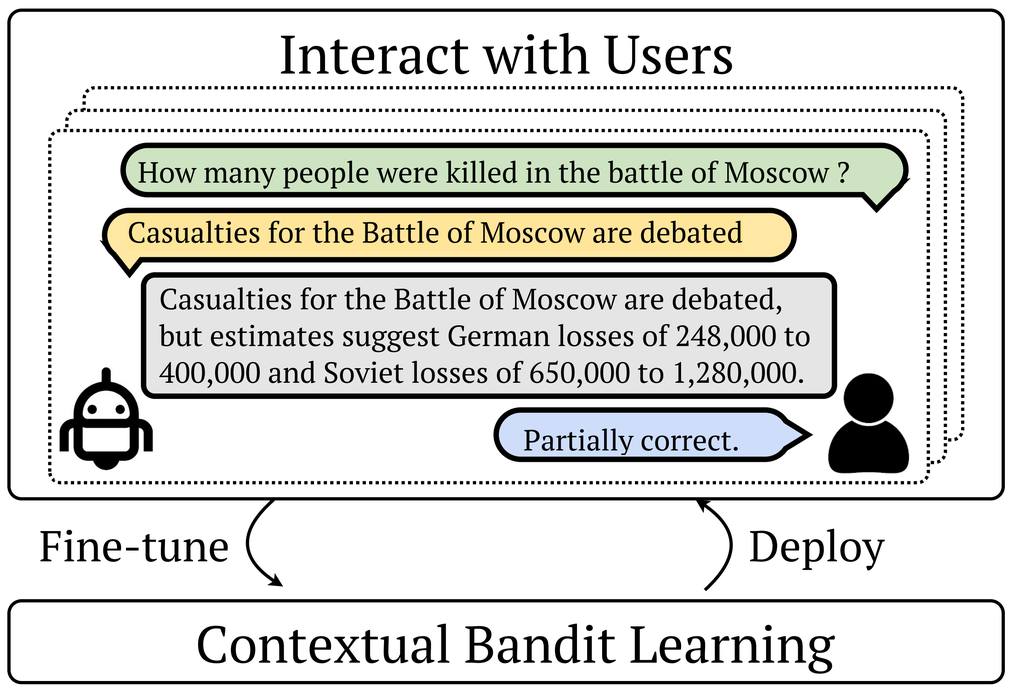
Code & Data

Research Question

How to improve NLP systems by learning from user feedback?

We present a user feedback study through bandit learning on extractive QA.

Feedback Setup



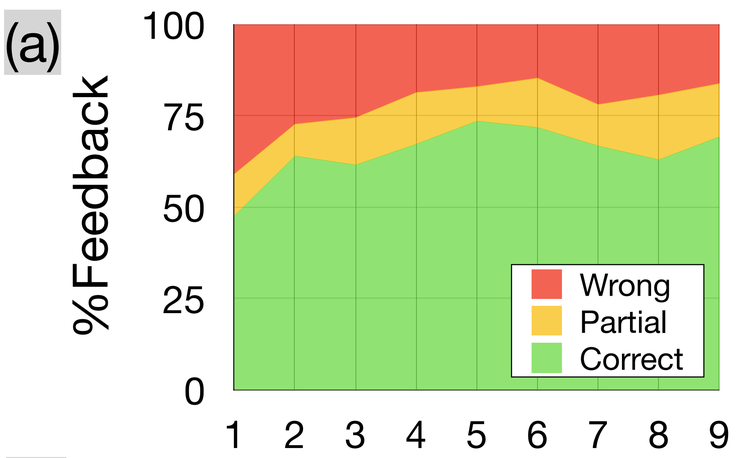
Settings

- 200 examples per round
- Hire MTurk workers
- Topics/Contexts from Wikipedia

Main Experiments

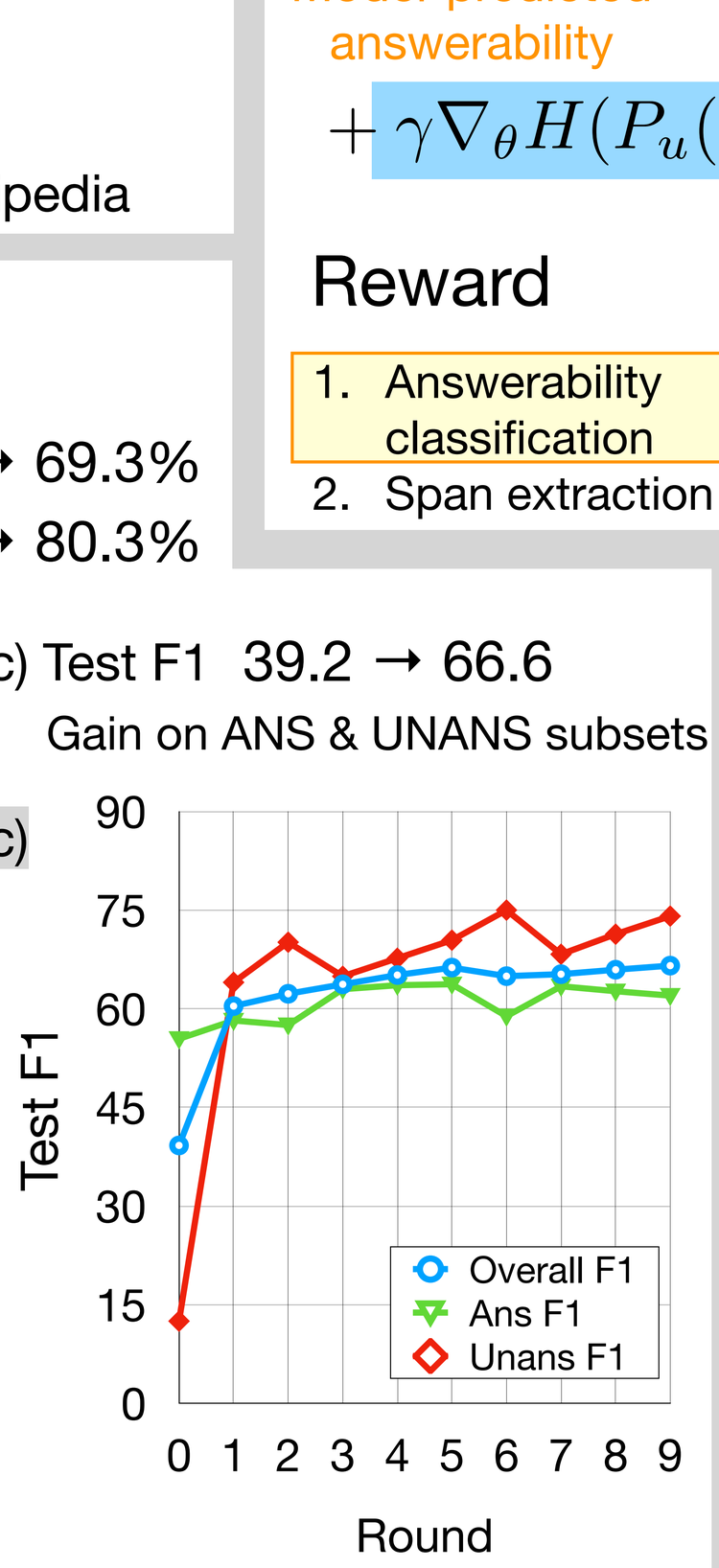
(a) Correct Feedback 47.7% → 69.3%

(b) Classification Acc 65.0% → 80.3%



(c) Test F1 39.2 → 66.6

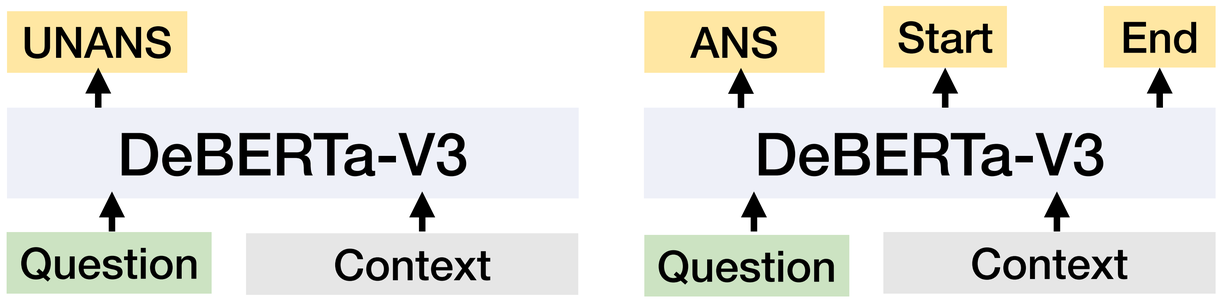
Gain on ANS & UNANS subsets



Task: Extractive QA

Question: Where does the name St Albans come from?
Answer: Alban
Context: <St Albans><Name> St Albans takes its name from the first British saint, Alban. The most elaborate version of his story, Bede's Ecclesiastical History of ...
Feedback: Partially Correct

Model: 3 classifiers, one for answerability, two for predicting start and end span



Algorithm

$$\alpha_1 r_1 \nabla_\theta \log \pi_\theta(\hat{u}|\bar{q},\bar{c}) + \alpha_2 r_2 \nabla_\theta \log \pi_\theta(\hat{y}|\bar{q},\bar{c}) + \gamma \nabla_\theta H(P_u(\cdot|\bar{q},\bar{c}))$$

$\frac{\pi_\theta(\hat{y}|\bar{q},\bar{c})}{\pi'_\theta(\hat{y}|\bar{q},\bar{c})}$ Inverse Propensity Score (debiasing reward)

Context, Question, QA model

Model-predicted answerability

Reward

Model-predicted span

entropy penalty

Reward

	Action	Feedback		
		Correct	**Partially Correct**	**Wrong**
1. Answerability classification	**UNANS**	1	0	-1
	ANS	1	1	0
2. Span extraction	**[i, j]**	1	0.5	-0.1

Setup Variants

- Domain Adaptation: NewsQA-trained model adapts to user distribution
- Performance degrades without answerability classifier