Allen Institute for AI

More Bang for Your Buck:
Natural Perturbations for Robust QA

Me

Daniel Khashabi
Allen Institute for AI

Tushar Khot
Allen Institute for AI

Ashish Sabharwal
Allen Institute for AI

Dataset Construction Pipeline

- Many NLP models remain data-hungry.
  - Large & rich datasets
- Dataset construction is often implemented as a repeated process of creating new instances by human annotators.

(Yonge Street) The Guinness Book of World Records no longer lists Yonge Street as the longest street in the world and has not chosen a replacement street, but cites the Pan-American Highway as the world's longest "motorable road'".

Human Annotators

Q: Is the "Yonge Street" the longest street in the world? ✗

- This can be a costly step and bottleneck for building stronger NLP models.

An Alternative Construction

We explore a slightly different dataset construction pipeline:

(Yonge Street) The Guinness Book of World Records no longer lists Yonge Street as the longest street in the world and has not chosen a replacement street, but cites the Pan-American Highway as the world's longest "motorable road'".

Q: *Is the "Yonge Street" the longest street in the world? (ans: ✗)*

↳ ***Q:*** *Was the "Yonge Street" the longest street in the world in the past? (ans: ✓)*

↳ ***Q:*** *Was the 'Yonge Street' the longest street in the world before 1980? (ans: unknown)*

natural perturbations != adversarial perturbations [Jia & Liang, 2017]

Creating natural perturbations are easier than writing new questions.

Conjecture: building datasets w/ natural perturbations are more **cost-efficient**.

Perturbing Boolean Questions

BoolQ
[Clark et al, 2019]

→ **Question perturbations** → **Question Verification** → BoolQ perturbations
collected on
Amazon Mechanical Turk

(1) Minimal changes
(2) Often flip the answer

Q: *Is the "Yonge Street" the longest street in the world? (ans: ✗)*

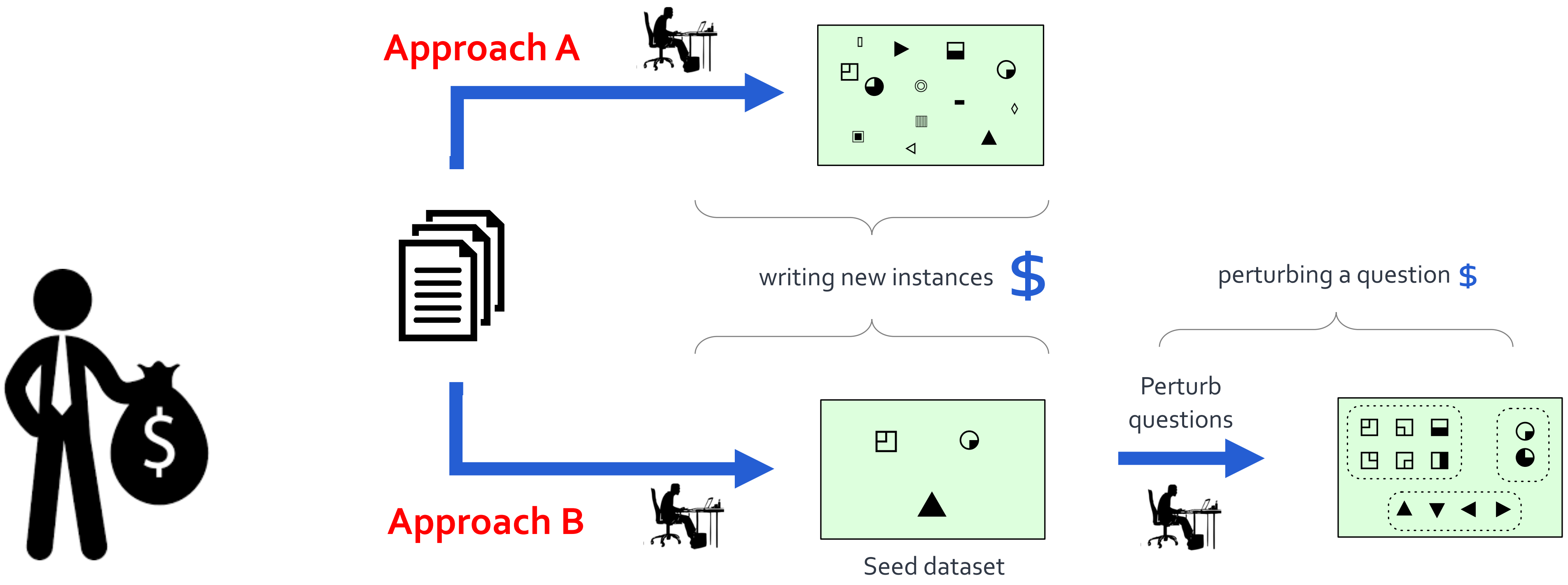
(Yonge Street) The Guinness Book of World Records no longer lists Yonge Street as the longest street in the world and has not chosen a replacement street, but cites the Pan-American Highway as the world's longest "motorable road"'.

Q: *Was the "Yonge Street" the longest street in the world in the past? (ans: ✓)*

Q: *Was the 'Yonge Street' the longest street in the world before 1980? (ans: unknown)*

Q: *Will the "Yonge Street" become the longest street in the world? (ans: ✗)*

Experimental Setup

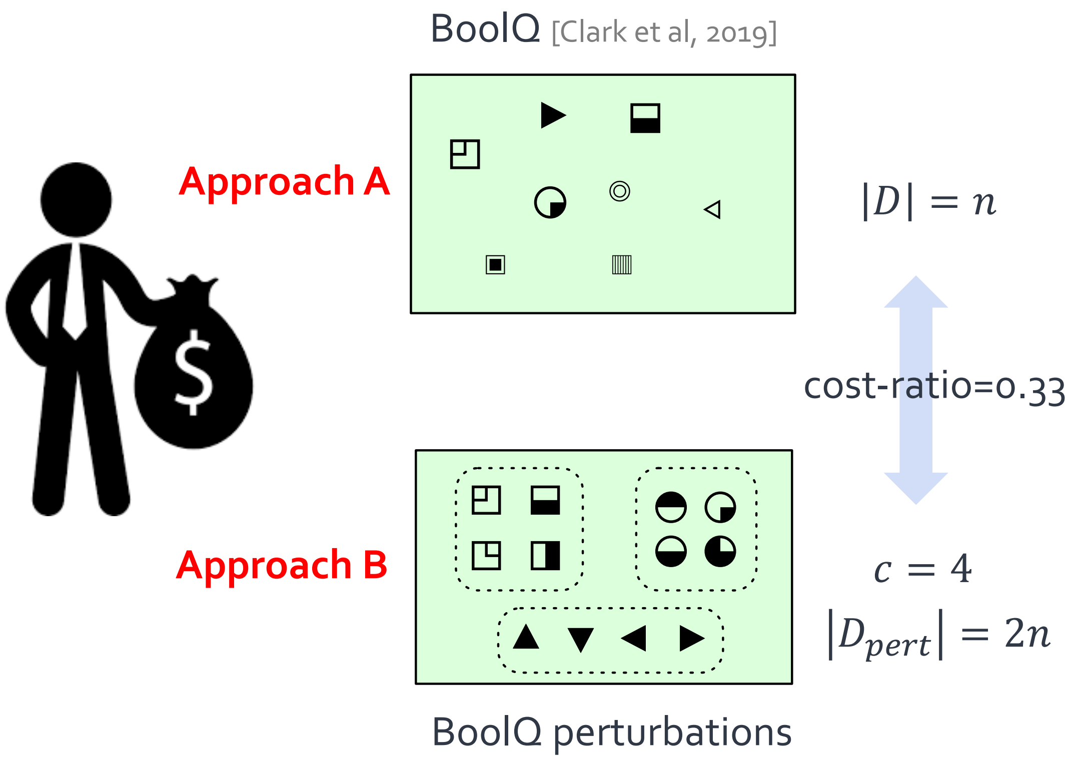


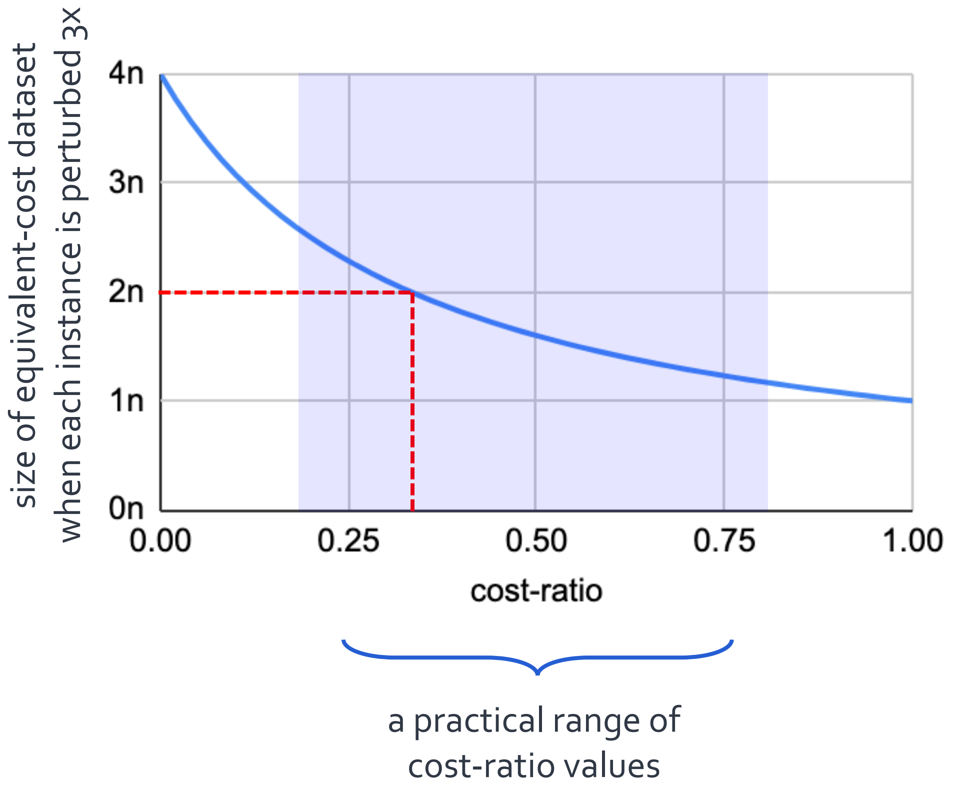
Intuition: perturbing questions is easier and cheaper than writing a question from scratch

$$\text{cost-ratio} = \frac{\text{perturbing a question } \$}{\text{writing a new question } \$}$$

Perturbation Size vs Cost-Ratio

- Equivalent-cost datasets, with varying cost-ratio.

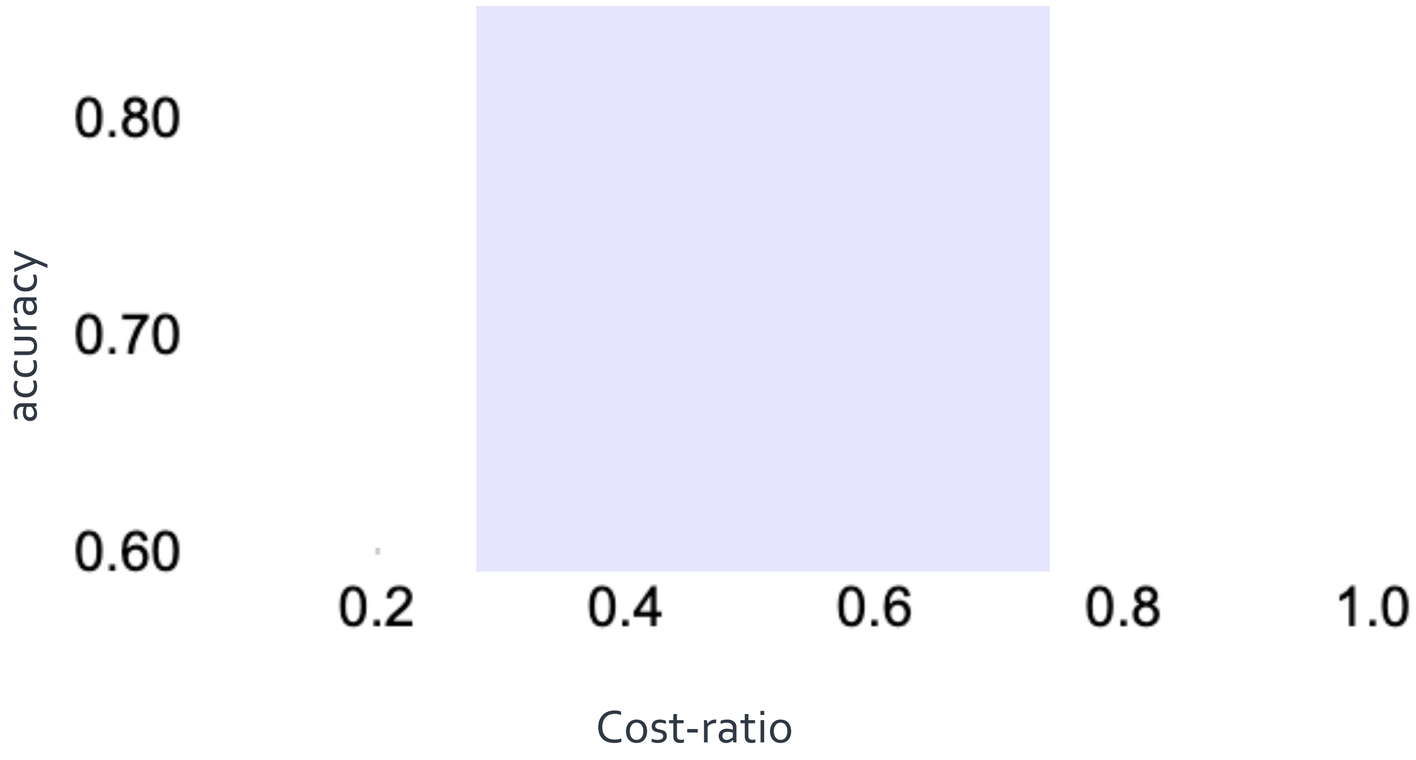




Experiment: Benefits as Function of Cost-Ratio

Evaluations on
BoolQ [Clark et al, 2019]

- Approach A (writing new instances)
- Approach B (each instance perturbed 3x)



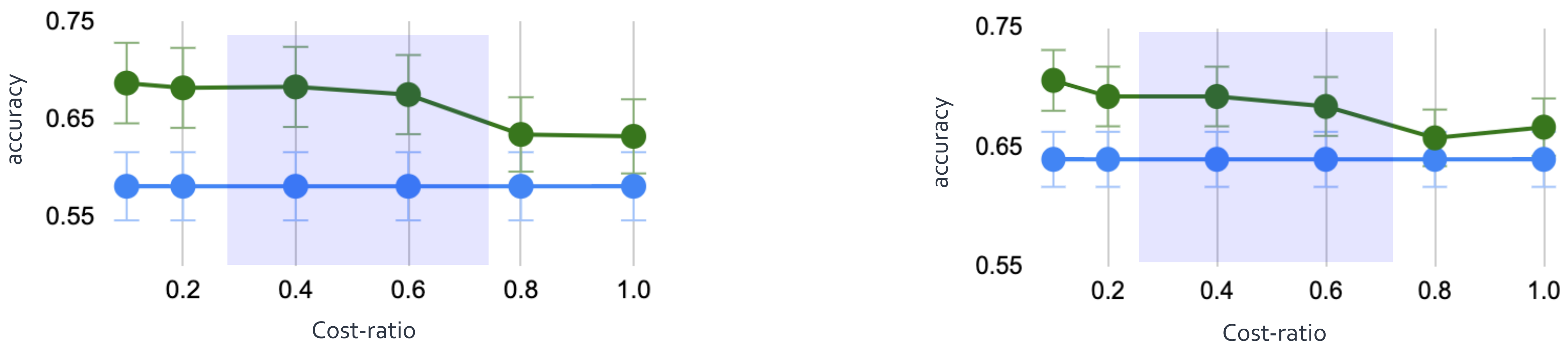
Observation:
Moderately cheap perturbations could result in more accurate models.

Experiment: Benefits as Function of Cost-Ratio

Approach A (writing new instances)
Approach B (each instance perturbed 3x)

Evaluations on
MultiRC [Khashabi et al, 2018]

Evaluations on
Contrast-Set of BoolQ [Gardner et al, 2020]



Observation:
Moderately cheap perturbations could result in better **generalization** to unseen data and and **robustness** to local perturbations.

Recent Work

- For reducing spurious associations [Kaushik et al., ICLR, 2020]
- NLI task [Huang et al., Workshop on Insights from Negative Results in NLP, 2020]
  - Do <u>not</u> observe any significant benefits from natural perturbations.

Summary

- An alternative approach for constructing training sets:
  - Expanding a seed set of examples via **human-authored perturbations**.

- When these perturbations are moderately cheaper, they result in gains.
  - Better generalization to unseen datasets.
  - Less sensitive to small changes in the input.

- Code & Data: `https://github.com/`**`allenai/natural-perturbations`**