

Supplementary Materials for ACL 2018 Paper: Semantically Equivalent Adversarial Rules for Debugging NLP Models

Interfaces for user studies

We present here screen shots of the interfaces for the experiments in Section 5 for the VQA dataset (sentiment analysis interfaces are similar, but do not have any images). All of these had accompanying instructions with examples, and/or a video tutorial.

Image

Original Question ?

Question: What color are the bird's beaks?

Answer: ● Orange.

AI probability of each option:

Orange.	0.77
Yellow.	0.16
Red.	0.02
Black.	0.05

Your question here ?

Which color is are the bird's beaks? Submit

Question: Which color is are the bird's beaks?

Answer: ● Orange.

This question did not change the A.I. answer

AI probability of each option:

Orange.	0.71
Yellow.	0.22
Red.	0.02
Black.	0.06

Status ?

6 questions left.

Legend:

✓ = changed the answer
✗ = didn't change the answer

Questions asked:

- ✗ What color are the bird's beaks?
- ✗ What colour are the bird's beaks?
- ✗ What is the color of the bird's beaks?
- ✗ Which color is are the bird's beaks?

Figure 1: Interface for condition **human** in Section 5.2. Subject is trying to create adversaries by modifying the input question.

Image

Evaluate similar questions

Original question:
- What kind of meat is on the boy's plate ?

Given the image on the left and the original question given above, which of the following questions is closer in meaning to the original question?

- What sort of meat is on the boy's plate ?
- What kind of meat is in the boy's plate ?
- What kind of meat are on the boy's plate ?
- What kind of meat is in the boy's dish ?
- What sort of meat is there on the boy's plate ?

Progress

Task 7 of 10.

Figure 2: Interface for condition **HSEA** in Section 5.2. Subject selects the SEA that is closest in meaning to the original question

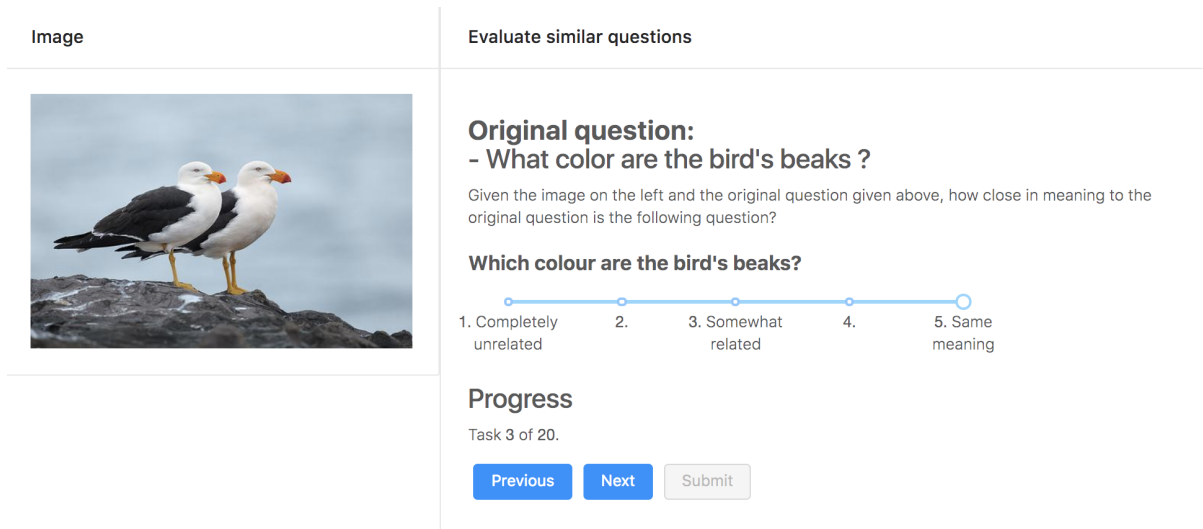


Figure 3: Interface for candidate evaluation in Section 5.2. Subject evaluates SEAs or human generated adversaries one at a time.

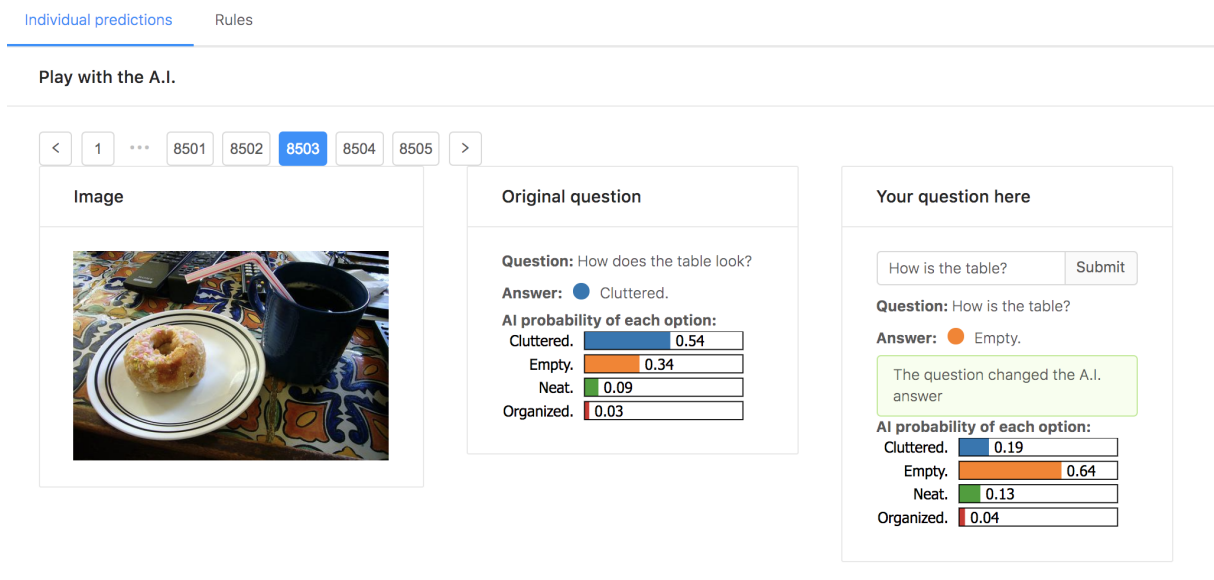


Figure 4: Interface for experts to play with the model, Section 5.3. Experts can get predictions for their own questions on validation images, and compare them to original predictions. Experts can move back and forth between this and the interface in Figure 5.

Try different rules

List of POS tags

Replace first instance of:

With:

Saved Rules

replace(Who is, Who's) x

replace(color, colour) x

Total Mistakes: 251

Mistakes if you save current rule: 549

Results

replace(What NOUN, Which NOUN)

Mistake examples (click images to see them in more detail)

< 1 2 3 4 > Compact




Image	Original	After rule
	Q: What color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.	Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.
	Q: What food item is above the burger ? Answer: a) Fries. b) Chips. c) Cole slaw. d) Ketchup.	Q: Which food item is above the burger ? Answer: a) Fries. b) Chips. c) Cole slaw. d) Ketchup.
	Q: What side of the court is the server playing on ? Answer: a) Left.	Q: Which side of the court is the server playing on ? Answer: a) Left.

Figure 5: Interface for experts to create and test rules , Section 5.3. Experts can see how many mistakes are induced by the current rule, and current saved rules (left), and see examples of mistakes produced by the rule with POS annotations (right).

Rules to evaluate

List of POS tags

Please look at the rule results on the right. The current rule is:

replace(What NOUN, Which NOUN)

Does the current rule induce a bug?

Progress

1 of 20.

Results

replace(What NOUN, Which NOUN)

Mistake examples

< 1 2 3 4 5 6 7 8 > Compact




Image	Original	After rule
	Q: What color are the pots ? Answer: a) Silver. b) Black. c) White. d) Gold.	Q: Which color are the pots ? Answer: a) Silver. b) Black. c) White. d) Gold.
	Q: What color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.	Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.
	Q: What animal is running in the background ? Answer: a) A dog. b) A horse. c) A llama. d) A kangaroo.	Q: Which animal is running in the background ? Answer: a) A dog. b) A horse. c) A llama. d) A kangaroo.

Figure 6: Interface for experts to evaluate SEARs. Experts were thoroughly instructed to only say “Yes” if a rule has semantic equivalence.