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.

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

Figure 2: Interface for condition **HSEA** in Section 5.2. Subject selects the SEA that is closest in meaning to the original question
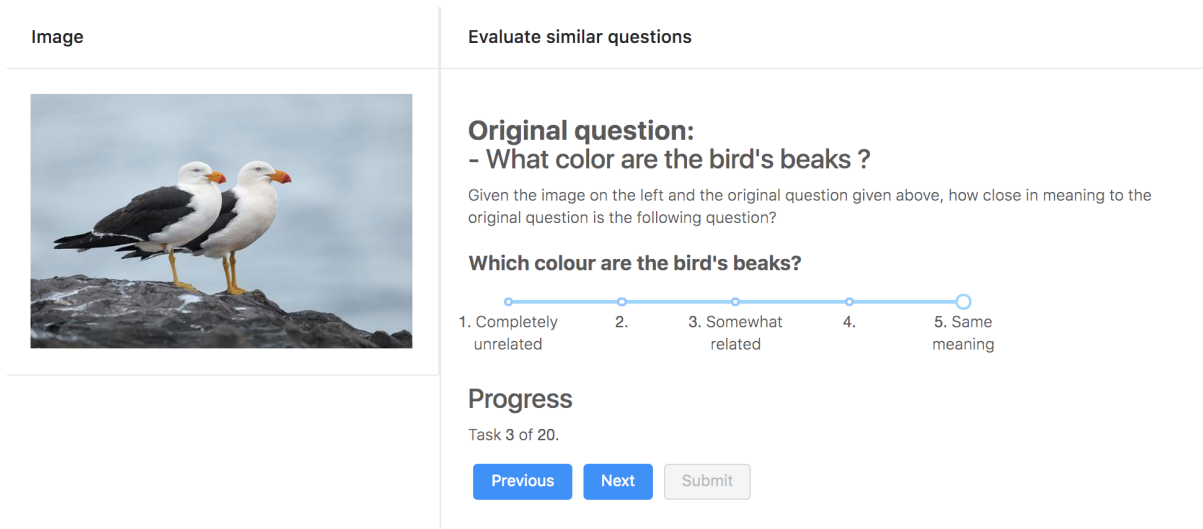
Figure 3: Interface for candidate evaluation in Section 5.2. Subject evaluates SEAs or human generated adversaries one at a time.
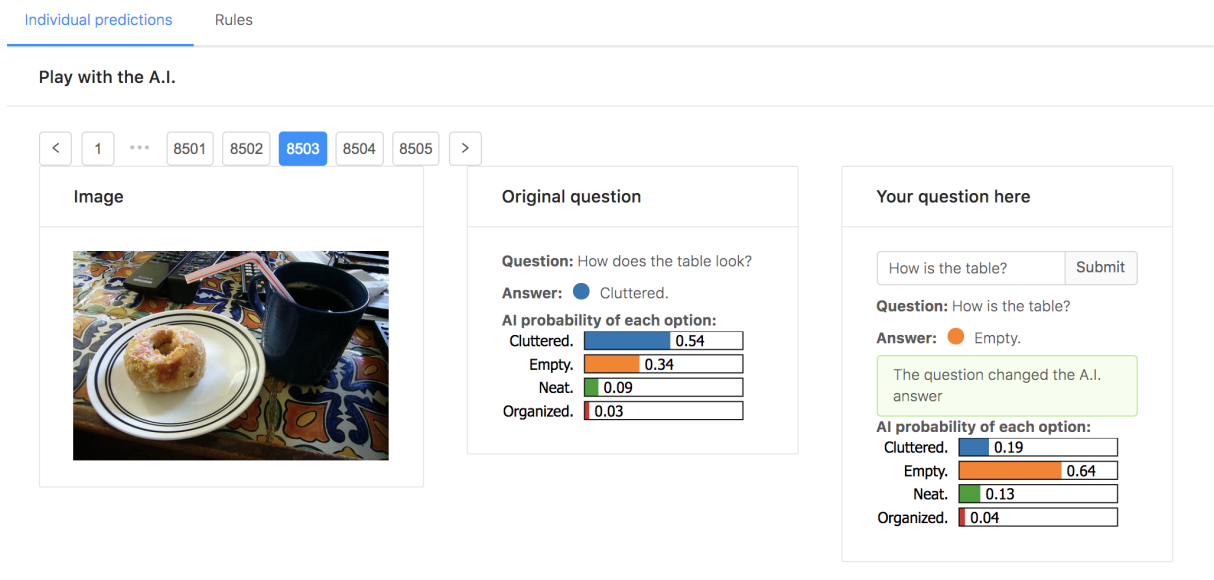

Figure 4: Interface for experts to play with the model, Section 5.3. Experts can get predictions for their own questions on validation images, and compare them to origial predictions. Experts can move back and forth between this and the interface in Figure5.

Try different rules

List of POS tags

Replace first instance of:

What NOUN

With:

Which NOUN   Submit

Saved Rules

replace(Who is, Who's) ×
replace(color, colour) ×

Total Mistakes   Mistakes if you save current rule

251   549

Finish

Results

replace(What NOUN, Which NOUN)

Save rule

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 ?   Q: Which color is the lampshade ?
Answer: Answer:
a) A light yellow.   a) A light yellow.
b) A bright red.   b) A bright red.
c) A subtle green.   c) A subtle green.
d) A vivid orange.   d) A vivid orange.

Q: What food item is above the burger ?   Q: Which food item is above the burger ?
Answer: Answer:
a) Fries.   a) Fries.
b) Chips.   b) Chips.
c) Cole slaw.   c) Cole slaw.
d) Ketchup.   d) Ketchup.

Q: What side of the court is the server playing on ?   Q: Which side of the court is the server playing on ?
Answer: Answer:
a) Left.   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?

No   Yes

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 ?   Q: Which color are the pots ?
Answer: Answer:
a) Silver.   a) Silver.
b) Black.   b) Black.
c) White.   c) White.
d) Gold.   d) Gold.

Q: What color is the lampshade ?   Q: Which color is the lampshade ?
Answer: Answer:
a) A light yellow.   a) A light yellow.
b) A bright red.   b) A bright red.
c) A subtle green.   c) A subtle green.
d) A vivid orange.   d) A vivid orange.

Q: What animal is running in the background ?   Q: Which animal is running in the background ?
Answer: Answer:
a) A dog.   a) A dog.
b) A horse.   b) A horse.
c) A llama.   c) A llama.
d) A kangaroo.   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.