## 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	0	Status	0
	Question: What color are the bird's beaks? Answer: Orange. Al probability of each option: Orange. 0.77 Yellow. 0.16 Red. 0.02 Black. 0.05		6 questions left. Legend: ✓ = changed the answer X = didn't change the answer Questions asked: X What color are the bird's beaks? X What is the color of the bird's beak	bird's beaks? bird's beaks?
	Your question here	0	× Which color is are the	ne bird's beaks?
	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			

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

Image	Evaluate similar questions	
	Original question:         - What color are the bird's beaks ?         Given the image on the left and the original question given above, how close in meaning to the original question is the following question?         Which colour are the bird's beaks?         1. Completely       2.       3. Somewhat       4.       5. Same meaning	
	Progress Task 3 of 20.	
	Previous Next Submit	

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

Individual predictions Rules		
Play with the A.I.		
< 1 ··· 8501 8502 8503 8504 8505	>	
Image	Original question	Your question here
	Question: How does the table look?         Answer:       Cluttered.         Al probability of each option:         Cluttered.       0.54         Empty.       0.34         Neat.       0.09         Organized.       0.03	How is the table? Submit Question: How is the table? Answer: Empty. The question changed the A.I. answer Al probability of each option: Cluttered. 0.19 Empty. 0.64 Neat. 0.13
		Organized. 0.04

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 Figure 5.

Try different rules	Results		
List of POS tags Replace first instance of:	replace(What NOUN, Which	NOUN)	
What NOUN With:	Save rule		
Which NOUN Submit	Mistake examples ( < 1 2 3 4 >	click images to see them in more detail)	
Saved Rules	Image Original	After rule	
replace(Who is, Who's) × replace(color, colour) × Total Mistakes Mistakes if you save current rule 251 549	G: What color is the lam Answer: (a) A light yellow, (b) A bright red, (c) A subtle green, (d) A vivid orange.	pshade ? Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. c) A subtle green. d) A vivid orange.	
Finish	G: What food item is ab C: What food item is ab C: What food Answer: (a) Fries) (b) Chips. (c) Cole slaw. (d) Ketchup.	ever the burger 2: Which food 2: Mich food 3: Fries. 4) Fries. 5: Cole slaw. 4) Ketchup.	
	Q: What side of the coupleying on ? Answer:	rt is the server Q: <mark>Which side</mark> of the court is the server playing on ?	

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	Results		
List of POS tags Please look at the rule results on the right. The current rule is:	replace( <mark>What NOUN</mark> , <mark>Which NOUN</mark> )		
replace(What NOUN, Which NOUN) Does the current rule induce a bug? No Yes	Mistake examples           <         1         2         3         4         5         6         7         8         >         Compact		
	Image Original	After rule	
Progress	<b>Q: What color</b> are the <b>Answer:</b> (a) Silver, b) Black. c) White. d) Gold.	pots ? Q: Which color Answer: a) Silver. b) Black. c) White. d) Gold.	
	<b>Q: What color</b> <b>Answer:</b> (a) A light yellow, b) A bright red. c) A subtle green. d) A vivid orange.	ampshade ? Q: Which color is the lampshade ? Answer: a) A light yellow. b) A bright red. (c) A subtle green. d) A vivid orange.	
	G: What animal is run Answer: (a) A dog. (b) A horse. (c) A llama. (d) A kangaroo.	ning in the background ? Q: Which animal Answer: a) A dog. (b) A horse. c) A liama. d) A kangaroo.	
	ou <sup>N</sup>	.0 <sup>114</sup>	

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