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.

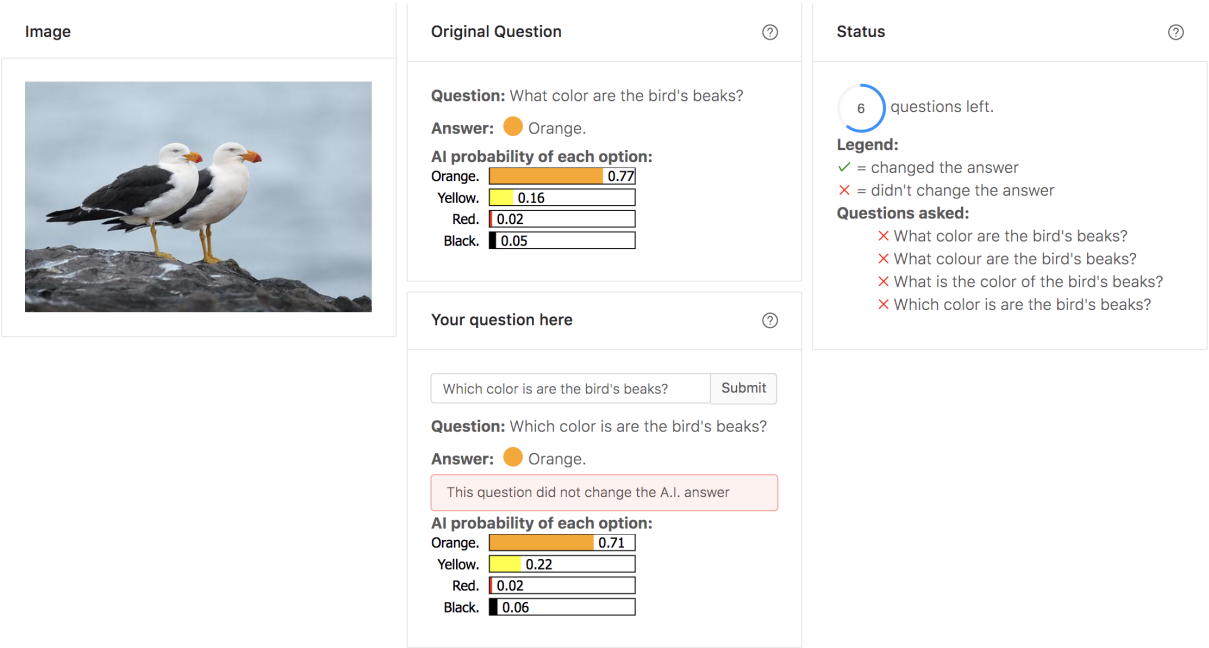

Figure 1: Interface for condition **human** in Section 5.2. Subject is trying to create adversaries by modifying the input question.

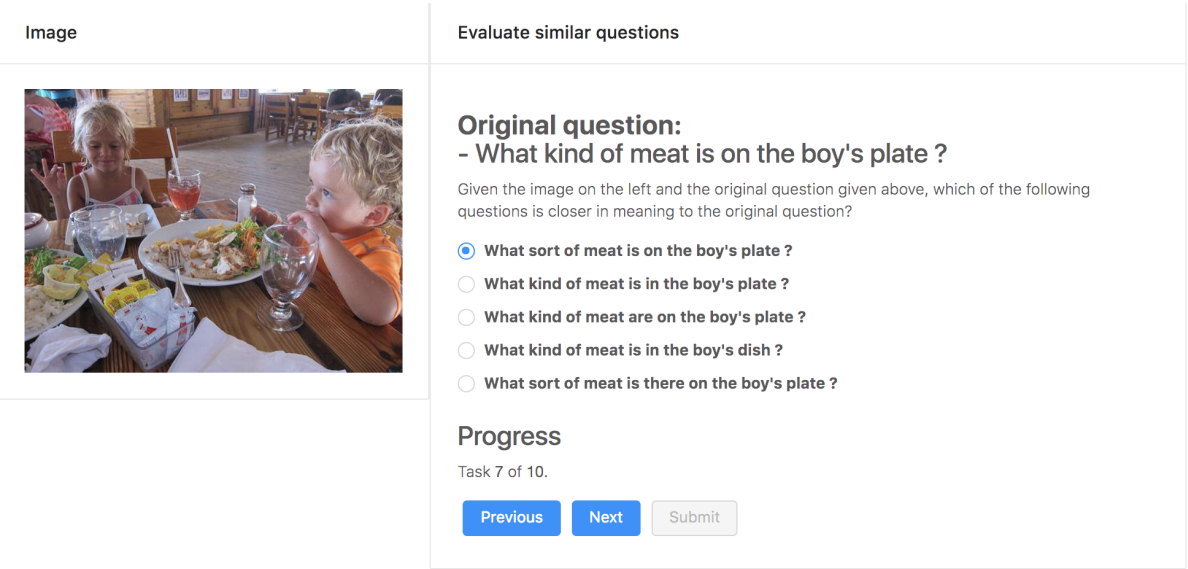

Figure 2: Interface for condition **HSEA** in Section 5.2. Subject selects the SEA that is closest in meaning to the original question

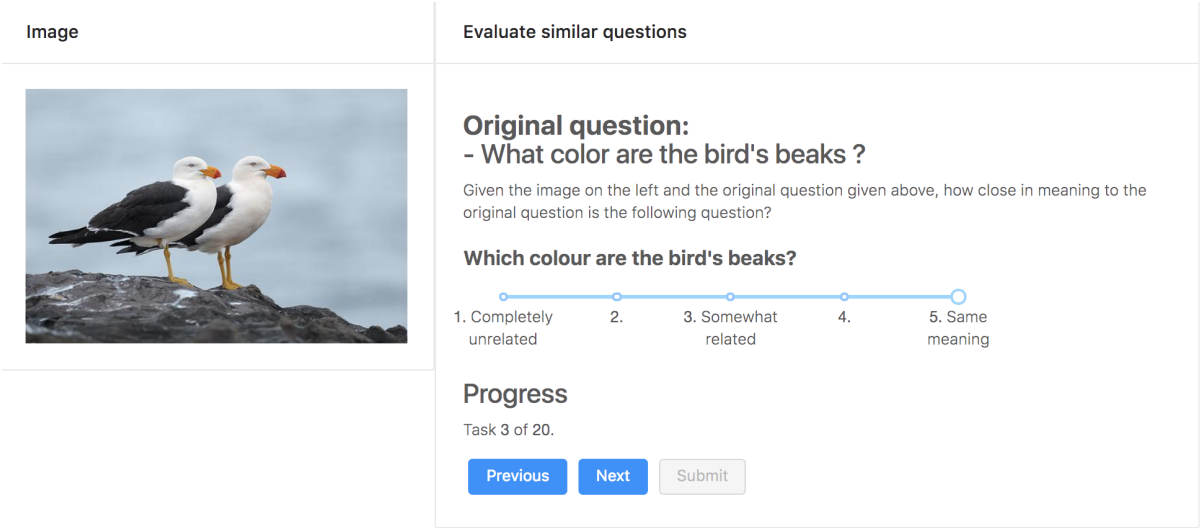

Figure 3: Interface for candidate evaluation in Section 5.2. Subject evaluates SEAs or human generated adversaries one at a time.

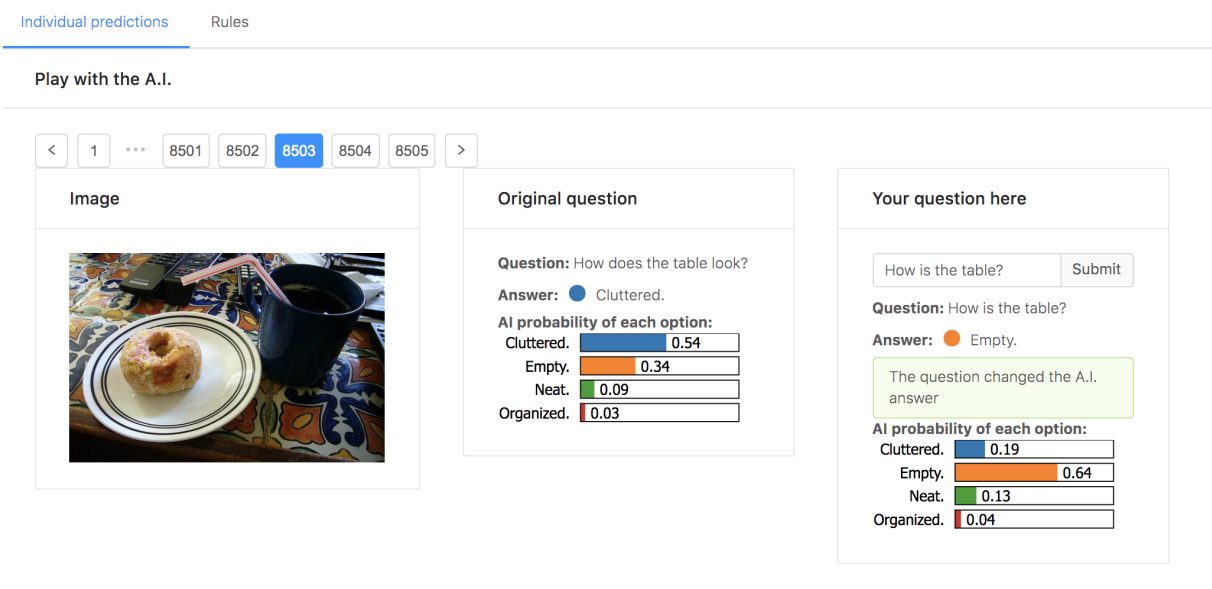

Figure 4: Interface for experts to play with the model, Section 5.3. Experts can get predictions for their own questions on validation images, and compare them to origial predictions. Experts can move back and forth between this and the interface in Figure5.

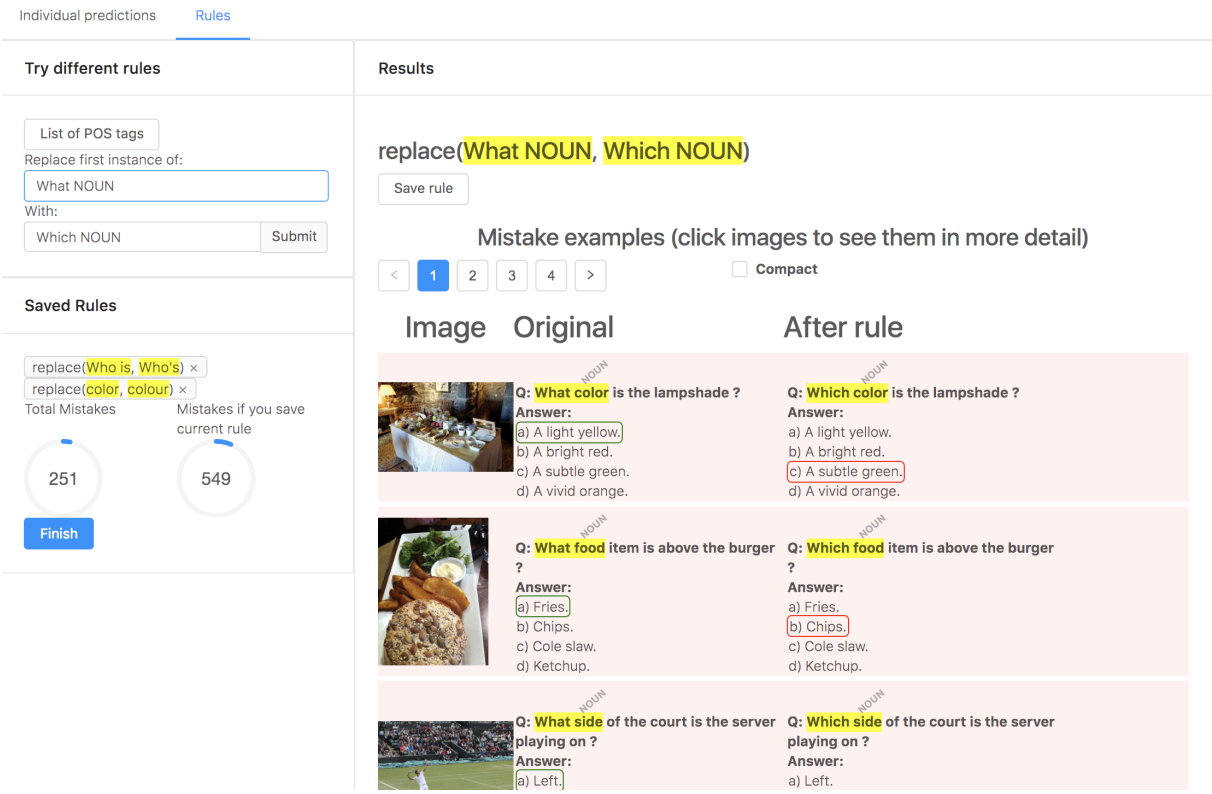

Figure 5: Interface for experts to create and test rules , Section 5.3. Experts can see how many mistakes are induced by the current rule, and current saved rules (left), and see examples of mistakes produced by the rule with POS annotations (right).

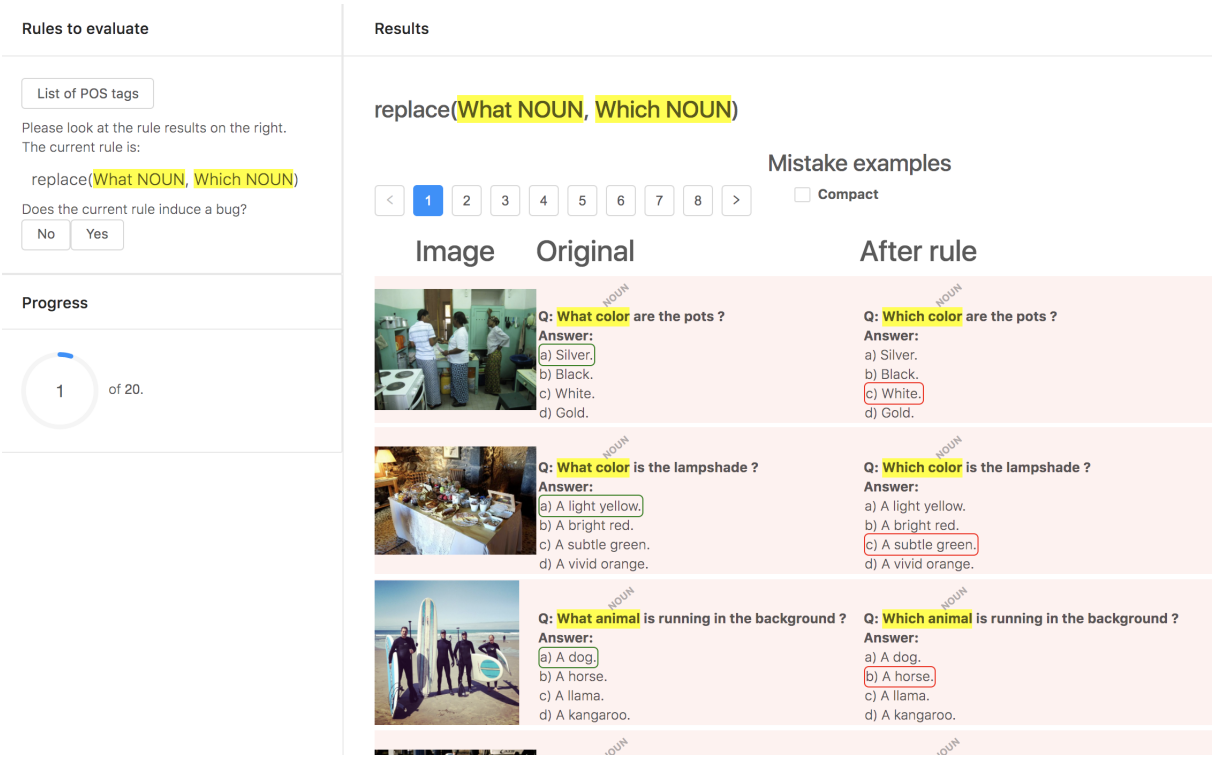

Figure 6: Interface for experts to evaluate SEARs. Experts were thoroughly instructed to only say "Yes" if a rule has semantic equivalence.