(B) RESUBMISSION: 1711
Original Action Editor: Marco Baroni

“Trick Me If You Can: Human-in-the-loop Generation of Adversarial Question Answering Examples”.

We have completed all mandatory revisions, (1) revising the writeup of the results sections and (2) adding hedges to the introduction about the specificity of our method.

We first discuss the mandatory revisions and then the points from the individual reviewers.

Mandatory Revisions:

Reviewers A and E expressed concern about the clarity of our results/experiments sections (sections 4 and 5). We have made revisions to make our results easier to follow, as well as clarify what research questions we are looking to answer (as suggested by the reviewers):

- 4.1 In the absence of an appropriate and quantifiable metric, we have removed the “fun” claim. Importantly, humans did not find the questions more difficult based on accuracy and speed.
- Added the motivation for why we used the Deep Averaging Network model and the Studio Ousia models at the beginning of Section 5. These models were included to test the transferability of the examples and thus the generalizability of the adversarial questions: does an example stump one model or **all** models?
- Figure 4 was awkwardly placed and was missing a clear reference. We moved Figure 4 to become Figure 6. We have also added a discussion which contrasts Figures 6 and 7 to highlight the differences between live human and model performance.
- We have added an introductory portion to Section 5 that outlines which models will be tested when.
- Sections 5.1-5.3 have motivation at the beginning of each section that explains why and what we are investigating in each subsection.
- Updated captions, legends, and titles for Figures 4--7 to increase clarity.
- Moved Section 5.4 to Section 6.1. Section 6 contains a detailed analysis of why the adversarial questions are harder for both IR and RNN models.
    - We have added an additional example of a “trick” that can be played on the RNN model in Section 6.2. In particular, novel clues can mislead the RNN model but not the IR model.

Reviewer D suggested our introduction should clarify that our paper and experiments are specific to Quizbowl despite the generalizability of our method. We made it clear that our use of human creativity to generate adversarial examples is only applied to Quizbowl in this work. Paragraph 3 of the intro has been changed to reflect this.

Furthermore, to make it clear what parts of our paper are generally applicable to NLP and which are specific to quizbowl:

- Section 2.1 describes our method and where it is generally applicable to NLP.
- Section 3-8 focuses specifically on Quizbowl, explaining how our method is specific to that dataset. We do not make claims that our work is applicable in any other setting.
- Section 9 further reiterates the necessary requirements to apply our framework to other tasks.

Reviewer A:

Reviewer A expressed confusion over whether the released logs will contain the adversarial questions as standalone data. We will include the final adversarial questions as standalone data. We will also include the complete edit history from the interface alongside the questions. In the final version of our paper, we will include links to external websites which will provide detailed information on the dataset.

Reviewer A wanted to see intuition for how Equation 1 models word removal. We have added the following clarification that it simulates *setting the vector to all zeros*:

“This simulates the change in prediction probability when a word’s embedding is set to the zero vector—i.e., approximating word removal—and is a common interpretation method for NLP (Ebrahimi et al., 2018; Wallace et al., 2018).”

Reviewer D:

Reviewer D wanted to see clarifications in the introduction that our work is specific to Quizbowl, we have modified the draft accordingly and addressed their concerns in the general response.

Reviewer E:

Reviewer E was curious why the adversarial questions were easier than normal questions for humans. The authors are sourced from well-trained Quizbowl writers---questions are of high quality. Human’s likely found the questions easier than normal questions because models have different failure modes than humans.

Figure 7: Studio Ousia’s model is not as similar to the attacked models as it appears based on the draft; we thank the reviewer for pointing out that we did not clearly describe the differences.

The Ousia model uses a knowledge graph (Freebase), an explicit entity type classifier, and customized word vectors (Wiki2Vec). We have modified the description and added *Appendix B* to provide further model details.

Reviewer E wanted to report average buzz position and accuracy for the live setting: this is shown in the current Figures 6 and 7. We have added a description to the text which compares the two figures. For the introduction statement (decreases as much as 40%), we meant a relative 40% decrease in accuracy (drop from 54.1% to 32.4%). We have reworded this to say "strong QA models only achieve 60% of their original performance", and made that result clear in Section 5.

We have addressed the remaining presentation improvements as suggested by the reviewers:

- 2.1 Clarified that our setup contrasts past human adversarial generation work (Ettinger et al. 2017) in that past work writes adversarial examples *independent* of a particular model (i.e., it's not model-driven or interactive like our work).
- 3.5 Corrected from "deep trivia" to "deep trivia knowledge"
- 7.1 Summarized Jordan Brownstein's main career accomplishment (national Quizbowl champion)

We thank the reviewers again for their helpful suggestions.