Anonymized Version of Original Decision Letter:

As TACL action editor for submission 1711, "Trick Me If You Can: Human-in-the-loop Generation of Adversarial Question Answering Examples", I am happy to tell you that I am accepting your paper subject (conditional) to your making specific revisions within two months.

LIST OF MANDATORY REVISIONS:

- as suggested by Reviewers A and E, revise the writeup of the results to i) tie the experiments more clearly to the the questions stated in the introductory parts of the paper; ii) discuss the results in more detail, along the lines suggested by A

- as suggested by (current) Reviewer D, add hedges to the introduction to clarify that the method is specific to quizbowl.

Generally, your revised version will be handled by the same action editor (me) and the same reviewers (if necessary) in making the final decision ---which, *if* all requested revisions are made, will be final acceptance.

You are allowed one to two extra pages of content to accommodate these revisions. To submit your revised version, follow the instructions in the "Revision and Resubmission Policy for TACL Submissions" section of the Author Guidelines at https://transacl.org/ojs/index.php/tacl/about/submissions#authorGuidelines .

Thank you for submitting to TACL, and I look forward to your revised version!

Marco Baroni
University of Trento and Facebook Artificial Intelligence Research
mbaroni@gmail.com

....THE REVIEWS....

Reviewer A:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

4. Understandable by most readers.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.:

3. Respectable: A nice research contribution that represents a notable extension of prior approaches or methodologies.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?:

4. Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:

• Authors should be informed of but not penalized for missing very recent and/or not widely known work.

• If a refereed version exists, authors should cite it in addition to or instead of the preprint.:

5. Precise and complete comparison with related work. Benefits and limitations are fully described and supported.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

4. They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

3. Potentially useful: Someone might find the new software useful for their work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

2. Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?:

4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.:

This paper has much improved since its first version, both in terms of writing and contribution. I particularly appreciate that annotation now takes place against both the IR and the neural system, identifying an asymmetry in the type of adversarial strategy needed to fool IR vs RNN models.

Although the authors have now added welcome statistics about the type of replacements performed by question writers, the analysis of that data is still on the short side (see Section 5.4). The main message I drew was the difference in behaviour between the IR and the RNN systems with respect to named entities. I understand that this explains the drop in accuracy in the IR model, as well as the fact that adversarial questions written for the RNN don't fool the IR system. I am still a little unsure which specific 'tricks' or phenomena can be played against the RNN itself. I would really like to see an analysis of this in the final paper.

Minor comments:
****

The questions were written against the IR and the RNN system, and then tested against the same two systems, as well as a DAN and the Ousia system. On first reading sections 5.1-5.3, it is a little hard to follow which model was tested in which round. It could perhaps be made clear in the introduction which types of models will be tested and how, and which are seen/unseen at test stage.

I'm not sure I follow equation 1 on l294. My (possibly incorrect) understanding is that the one-hot vector w_i picks out one of the words in the sequence w and we are computing the derivative of the classifier (approximated as a linear function) with respect to w_i. I don't get how using w_i models the removal of a word from the sequence. The cited work in that section is probably helpful in that respect but it would be nice if a better intuition of the technique was given in the text.

I don't fully understand whether the released logs will contain the adversarial questions as standalone data.

REVIEWER CONFIDENCE:
4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Reviewer D:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:
5. Very clear.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.:
4. Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?:

4. Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:
• Authors should be informed of but not penalized for missing very recent and/or not widely known work.
• If a refereed version exists, authors should cite it in addition to or instead of the preprint.:

5. Precise and complete comparison with related work. Benefits and limitations are fully described and supported.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

4. Some of the ideas or results will substantially help other people's ongoing research.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

3. They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

1. No usable software released.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

4. Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?:

5. Strong: I'd like to see it accepted; it will be one of the better papers in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.:

I reviewed a previous version of this work (I am "reviewer A"). I think the current revision addresses all of my concerns, as well as (to my judgement) also the concerns of the other reviewers. I agree with the authors' assessment that the current version is much stronger than the previous one, and I'd recommend it for publication.

A small remaining issue is that I still feel that the introduction narrative is making claims about a general method while in fact the paper is still very much specific to quizbowl. To be clear: the focus on quizbowl is not a problem, but the claim to generality in the intro is a bit off-putting to me. But I can also see it as being a taste-issue, and would not argue strongly for its removal. I do not think there is a real danger that a reader will be mislead by that.

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Reviewer E:

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

3. Mostly understandable to me (a qualified reviewer) with some effort.

INNOVATIVENESS: How original is the approach? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?

Note that a paper can score high for innovativeness even if its impact will be limited.:

4. Creative: An intriguing problem, technique, or approach that is substantially different from previous research.

SOUNDNESS/CORRECTNESS: First, is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments and are the results of the experiments correctly interpreted?:

4. Generally solid work, although there are some aspects of the approach or evaluation I am not sure about.

RELATED WORK: Does the submission make clear where the presented system sits with respect to existing literature? Are the references adequate?

Note that the existing literature includes preprints, but in the case of preprints:
• Authors should be informed of but not penalized for missing very recent and/or not widely known work.
• If a refereed version exists, authors should cite it in addition to or instead of the preprint.:

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included. Discussion of benefits and limitations is acceptable but not enlightening.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

4. Some of the ideas or results will substantially help other people's ongoing research.

REPLICABILITY: Will members of the ACL community be able to reproduce or verify the results in this paper?:

4. They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

3. Potentially useful: Someone might find the new software useful for their work.

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

3. Potentially useful: Someone might find the new datasets useful for their work.

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?:

4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.:

SUMMARY

Quiz Bowl is a competition that tests players on a variety of subjects, which was used as the basis for the NIPS 2017 Human-Computer Question Answering Competition. The manuscript proposes an adversarial interface for authoring Quiz Bowl questions, that introspects IR and neural models to highlight the words they use to identify answers. Human authors can then experiment with different phrasing.

Interestingly, a human competition over questions authored using the interface are easier. On average, competitors:

- buzz after 60% of the question has been read, compared to 72% for Baseline;
- achieve an accuracy of 90%, compared to 84% for baseline.

By contrast, the Studio Ousia system that won the NIPS competition (Yamada et al., 2018) struggles on interface questions. Playing against a comparable human team, it loses 300-30, compared to winning 475-200 at the NIPS competition.

The manuscript suggests that the proposed human adversarial evaluation is novel. Instead of hypothesising phenomena and using humans to verify output, humans authors have complete freedom to rephrase questions in whatever way the like. The manuscript categorises and counts resulting phenomena across 100 questions, with the most common phenomena being paraphrasing, introducing ambiguity, and multi-step reasoning.

Overall, a very interesting submission, thank you. I did have a bit of a hard time following the narrative in places, in particular in the results sections. More detailed comments follow.

QUESTIONS/COMMENTS

* I had a bit of a hard time following the experimental results in Sections 4 and 5. Suggest an editorial pass that ties these back to the results summary abstract/introduction. I think a bit of hand holding in each subsection would help, e.g., explicitly stating research questions, using one figure/table per question, explaining how to interpret figures/tables, summarising conclusions from figures/tables.

* Figure 4: I didn't see a reference to this Figure so wasn't sure which section it corresponds to. Suggest keeping Figures/Tables as close to their first mention as possible, and making sure there is a reference in the text that includes description and discussion.

* Section 4.1: Why is human performance better on questions authored using the interface? Does it have to do with the authors, e.g., their Quiz Bowl experience/expertise? Or does it have to do with the models used in the interface, e.g., having different characteristic failure modes than humans? Who were the adversarial authors and how were they sourced?

* Can you report average buzz position and accuracy for the live humans vs computers setting? I couldn't figure out which result that ties back to the statement in the introduction that "the accuracy of strong QA models decreases as much as 40%."

* In Figure 7: "The performance of the state-of-the-art Ousia model degrades on the adversarially-authored questions despite never being targeted by the adversarial authors." While the Ousia model wasn't targeted directly in the interface, the models used are very similar. Isn't this kind of the whole point, i.e., to rephrase questions so they confound models using IR and neural Quiz Bowl components?

OTHER QUESTIONS/COMMENTS

* Section 2.1: "which contrasts past human adversarial generation" —Contrasts how?

* Figure 3: Possible to make this a bit bigger / easier to read?

* Section 3.5" "they have deep trivia and craft questions" — Something seems to have gone wrong here?

* Section 3.5: "rather, questions are difficult because of their semantic content" — Isn't the key point that the approach facilitates semantic as well as syntactic rephrasing?

* Section 4.1: "The next test of validity is whether humans find the questions fun and challenging" — This suggests to me some kind of human feedback, e.g., Likert ratings for fun. However, I only saw results for accuracy and answer speed. Suggest rephrasing.

* Section 5: Why is the deep averaging network included in results? I don't think it is mentioned before this point, so could use introduction and motivation. If used here, why not used in UI as well?

* Figures are difficult to read in black and white.

* Section 7.1: "another interviewee was Jordan Brownstein, one of the greatest players of all time" — This is a bit informal. Is it possible to summarise specific career achievements in a few words instead?

REVIEWER CONFIDENCE:

4. Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.