==========
NeuralGen 2019 Reviews for Submission #32
==========

Title: On conducting better validation studies of automatic metrics in natural language generation evaluation
Authors: Johnny Wei
==========
REVIEWER #1
==========

Reviewer's Scores

Appropriateness (1-5): 5
Clarity (1-5): 3
Originality / Innovativeness (1-5): 2
Soundness / Correctness (1-5): 4
Meaningful Comparison (1-5): 4
Thoroughness (1-5): 3
Impact of Ideas or Results (1-5): 3
Recommendation (1-5): 3
Reviewer Confidence (1-5): 2

Detailed Comments

Key Contributions:

This paper discusses the issue of evaluating automatic evaluation metrics (AEMs). It discusses how AEMs have been evaluated within the machine translation community, presents considerations and advice from literature about how to evaluate an AEM, and presents some examples of evaluating systems from WMT'17.

Strengths:

This paper is basically a best-practices guide to conducting an evaluation of an AEM. As such, this could be valuable to the community, if people are unsure how to go about doing this. It discusses a number of considerations that validations studies should consider, such as system or segment level evaluation, some pros and cons of different measure of correlation, and a basic recommendation of which statistical test to use. It contains extensive citations to the literature for those who want to know more.

Weaknesses:

This paper feels like a bit of a missed opportunity. There has been a proliferation of metrics, and it might be useful to have a coherent approach to determining which of these is best, and how they could be improved.

This paper, however, is somewhat unambitious with respect to this larger goal. The majority of the paper seems to be summarizing points that have been made elsewhere (with proper citations, of course), including figures, so it is not clear if there is any novelty. Although it points to some limitations of existing metrics, it seems like it could go much farther with critiquing existing AEMs. Moreover, the paper hedges somewhat with respect to what we want from an AEM. Most of the emphasis is on linear correlation between a metric's values and human judgements, but the paper also notes that it is important to distinguish cases where, for example, a metric can distinguish between good and mediocre, but not mediocre and bad. The fact that this is an issue suggests!
that we should be working with a non-linear relationship between AEM values and human judgements. The reasons for using Pearson correlation given do not seem convincing: significance tests are possible for other measure as well. Why not go farther, and try to map AEM values directly to human judgements on a common scale using a simple non-linear function? We could then evaluate in terms of mean-squared error on held out data, or some similar metric.

Although the writing is fairly clear, it is somewhat vague in places (e.g. "the choice of question is nearly art [sic] and considerations may be philosophical."). Similarly, some of the advice seems unjustified (e.g. bin the DA scores into good, average and bad). Lastly, the existing conclusions section does not offer conclusions, but rather some very vague speculations ("The authors believe... This metric might ... "). My sense is that this paper might be most profitably written as a succinct short paper which provides concrete advice and pointers to the literature. It would probably have much more influence in that form.

==========

REVIEWER #2

==========

Reviewer's Scores

Appropriateness (1-5): 5
Clarity (1-5): 5
Originality / Innovativeness (1-5): 3
Soundness / Correctness (1-5): 5
Meaningful Comparison (1-5): 5
Thoroughness (1-5): 4
Impact of Ideas or Results (1-5): 4
Recommendation (1-5): 5
Reviewer Confidence (1-5): 4

Detailed Comments

Key Contributions:
* The authors provide a prescriptive guideline for evaluating evaluation metrics by synthesizing observations across a number of related work.
* The authors analyze the metrics proposed in WMT '17 as an example of using their guidelines.

Strengths:
* The authors have synthesized a lot of information to produce their prescriptions: there has been a large body of work in comparing automatic metrics and having a single place to read this is useful.

Weaknesses:
* I think the impact of this paper could be significantly larger if the authors included more interpretation on the results, e.g. the Kendall correlations.

Questions for Authors

I think the numbers mentioned in section 4.4, line 697, should have been 76.7%?
