The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

ACL-IJCNLP 2021

Author Response

Title: ChatMatch: Evaluating Chatbots by Autonomous Chat Tournaments
Authors: Ruolan Yang, Zitong Li and Kenny Zhu

Instructions

The author response period has begun. The reviews for your submission are displayed on this page. If you want to respond to the points raised in the reviews, you may do so in the boxes provided below.

Please note: *you are not obligated to respond to the reviews*.

For reference, you may see the review form that reviewers used to evaluate your submission. If you do not see some of the filled-in fields in the reviews below, it means that they were intended to be seen only by the committee. See the review form HERE.

Review #1

The core review

This paper presents a new automatic evaluation framework called ChatMatch. The compare two chatbots based on three-level rule, i.e., repeation, contradiction and coherence with previous concept. The results show a good correlation with human judges. The model is too simple, and the writing is also not professional.

Reasons to Accept

N/A

Reasons to Reject

(1)The model is too simple, only using three-level rule, i.e., repeation, contradiction and coherence with previous concept. I think it is not enough for evaluating the quality of the generation from the chatbots, because there are many complex situations. (2)The algorithm of repetition detection and inconsistency detection only use the similarity of two sentences, which is not enough for measure the consistency and repetition. (3)They could use the bert-score as the state-of-the art baseline for comparison.

Overall Recommendation: 1.5

Review #2

The core review

This paper describes and proposes a novel idea on evaluating chatbots. ChatMatch is technically an evaluation framework that competes chatbot against another chatbot in the tournament-like competition. Later, judgement will take place to observe and assess the bot conversations through predefined metrics/rules such as repetition, inconsistency, and memorization.

Author manages to introduce the proposed approach clearly, the details in the paper are adequate enough to reproduce, and the experimental results are technically convincing. The only contribution of this paper was a novel idea of tournament-like scoring to chatbot evaluation. Overall this paper is well written and easy to understand.

However there are some parts where this paper failed to address. This paper doesn’t address reproducibility consistency issues at all. Which is quite an important issue in dialog system evaluation tasks. Given the proposed method, obviously anyone can get different results or score with different bots combination. How can it become a justifiable evaluation system if every evaluation iteration results are different? There is no clear reason why the authors pick 3 kinds of manual metrics (repetition, inconsistency, and memory ability) as a base algorithm for the scoring scheme. The reviewer feels that more explanation on this is necessary and could benefit the reader.

Reasons to Accept

This paper proposed a novel suggestion on chatbot-evaluation scheme. This paper can give an interesting perspective on dialog evaluation tasks to the dialog community. The proposed idea can be a fair evaluation system for a chatbot competition. On the positive note, hypothetically, a centralized dialog evaluation system might benefit from this proposal.

Reasons to Reject

Dialog evaluation task is a highly saturated labor. An evaluation task is meant to be easy to reproduce and give a consistent result, which is very far from what the author proposed here. While the proposal is interesting and original, there is no clear discussion on how the community could benefit from this idea. This is a very major flaw.

Overall Recommendation: 2.5

Missing References

There is no clear justification on error taxonomies applied on the paper. The reviewer suggests that the author has more justified metrics, it can be based on this paper. https://www.aclweb.org/anthology/W15-4611.pd

Typos, Grammar, Style, and Presentation Improvements

The reviewer feels that section 4 (related works) can be moved forward after the introduction section. That way readers can have a nice perspective on existing works, and how the presented ideas adds to that.

Review #3

The core review

This paper introduces a method for automatic chatbot evaluation dubbed ChatMatch where different chatbots chat with each other. They subsequently rate each bot in these conversations as using repetition, inconsistency, and long-term conversational history metrics. Each time the bots interact twice alternating which one starts at the second turn after a seed prompt to start the conversation. Finally, they match bots in a tournament style and generate overall rankings between the chatbots. The authors compare their method against perplexity, token accuracy, and coherence automatic metrics as well as the spot-the-bot human evaluation. Furthermore, they enlist undergraduate students to create a gold standard ranking of the models using interactive evaluation. Overall, they find strong correlation with human evaluation while having an extremely fast automatic metric.

I really like the idea of a joint metric to assess the features. I think that the modules are simple and an interesting non-parametric way of evaluating the chatbots. Each of the three modules could be replaced with separate general units where these could be further improved upon.

In my opinion the main insurmountable problem of this paper is the lack of state-of-the-art methods (e.g. Blender 2.7B, DialoGPT fine-tuned on BST, Plato (Bao et al 2020)). While I like the idea, I think the authors have a lot of room to improve upon in exploring the metrics as well as human evaluation. I believe that their goal of precision(higher \tau) is in conflict with accuracy (correlation with human judgement).

Reasons to Accept

There is a huge need for automatic metrics for chatbot evaluation. While we've seen a lot of improvement, I think that this is a unique approach to the problem that shows real promise.

Reasons to Reject

I like the overall approach taken by the authors, but I think that there is more work required before publication. 1. as stated in the weaknesses, there should be more state-of-the-art chatbots (e.g. Blender 2.7B, DialoGPT fine-tuned on BST, Plato (Bao et al 2020)) and 2. a careful assessment of which modules are best for each for consistency/repetition/memory. The authors favor precision over accuracy meaning that their regression weight shows lower consistency and so it is not preferred.

Overall Recommendation: 2.5

Questions for the Author(s)

In table 4 the average \tau is 0.35 how many seed sentences are below that in value?

Why is high precision (\tau) more important than accuracy(correlation with human judgment)?

What do the last 5 exchanges in the conversation look like? From figure 3 it looks like there is a lot of benefit to having the chatbots talk for really long sequences (well beyond human).

Typos, Grammar, Style, and Presentation Improvements

It would be good to have one table that showed all of the rankings clearly side-by-side.

move figure 1 to the top of the page.

line 546 ",We" line 630 space before "(" line 638 rephrase "more potent"

Submit Response to Reviewers

Use the following boxes to enter your response to the reviews. Please limit the total amount of words in your comments to 1000 words (longer responses will not be accepted by the system).

Response to Review #1:

Response to Review #2:

Response to Review #3:

General Response to Reviewers:

Response to Chairs

Use this textbox to contact the chairs (including area chair and senior area chairs) directly only when there are serious issues regarding the reviews. Such issues can include reviewers who grossly misunderstood the submission, or have made unfair comparisons or requests in their reviews. Most submissions should not need to use this facility.

Submit

START Conference Manager (V2.61.0 - Rev. 6290)