Review #1

What is this paper about and what contributions does it make?

This paper proposed FocusL, a novel training objective for faithful knowledge-grounded dialogue response generation. Compared to traditional cross-entropy training objectives, FocusL scales the original logit (model prediction) at each time step with the relevance (similarity) between the corresponding token and the grounding knowledge (please see line 283). The proposed method achieved significantly better faithfulness and fluency results with automatic metrics on two datasets (WoW and FaithDial). It is also preferred by human evaluators based on the results from a human study on the WoW dataset.

Reasons to accept

1. The paper focuses on an important issue of hallucinations in dialogue response generation. The proposed method is simple yet effective, showing a promising future research direction.
2. The introduced experimental study is compressive.
3. The paper is overall well presented.
4. This paper performed statistical significance tests to show the effectiveness of the proposed method.

Reasons to reject

1. Section 3. The notations in section 3.1 are confusing. For example, "C_1" in line 173 is an utterance. However, in line 175, "k_m" is a token instead. Please use a different font or type of characters to distinguish them. In line 178, it is unclear why "T" is capitalised. These inconsistencies of the notations across this paper hinder the reader from understanding the proposed method.
2. In addition, the explanation of the proposed method is not crystal clear. Please also answer my question B, C, E, F, and G.
3. The concept of focus is a bit unclear to me. This focus is very similar to attention. The improvement from the proposed method could also be achieved by other means, which may be simpler. Please see my question A and D.
4. I also have some confusion about the experimental results and the findings derived from them. Could you answer my question E and F?
5. Please also note that some comments in the submitted codes may reveal your identity. Please be careful in the future.

Questions for the Author(s)

Question A: I am unfamiliar with the concept of "focus". Is this concept proposed in this paper, or are there any previous works? Could you explain the difference between the concepts of focus and attention?

Question B: Line 272. Could you explain more about the intuition behind this proposed mapping formula? Or could you give me an example where this similar method has been applied to other problems?

Question C: Line 283. Is "w_a" a vector or scalar?

Question D: Line 306. Have you tried a different value of p? I wonder if decreasing the value of p will achieve a similar effect as the proposed method.

Question E: Line 432 - 433. Have you trained your model on the WOW dataset? If so, could you explain why you called this "out-of-domain"?

Question F: Tables 1, 2, and 3. These tables mention the results of the statistical significance test, which is very nice to include. Could you introduce more about how those results are calculated?

Question G: Line 476. Could you explain more about why the weight distribution with the threshold is also named CW? I am sorry, but this name seems a little bit unclear to me.

Typos, Grammar, Style, and Presentation Improvements

1. Line 122. The word "dialogue" is also spelt as "dialog" in the paper.
2. Line 141. "Hallucinations in Text Generation".
3. Line 401 and 432, "We" -> "we".
4. Could you try to enlarge Figure 5?

Soundness: 4
Excitement (Long paper): 3.5
Reviewer Confidence: 4
Recommendation for Best Paper Award: No
Reproducibility: 5
Ethical Concerns: No

Review #2

What is this paper about and what contributions does it make?

This paper focuses on the important problem of reducing hallucinations in open domain conversation system responses, which makes for more interesting and engaging chatbots. The main technical contributions are the updated loss function, which gives additional weight to content words considered relevant based on matching content in a knowledge graph, and a set of experiments detailing the impact of the method. The idea of weighting the words differently is not new (I included a very relevant reference below) but the approach detailed here is still relevant.

Reasons to accept

The method is described fairly clearly and with sufficient detail to reproduce the implementation. The experiments are fairly well set up and thorough, with informative ablation studies. The method is using at least a number of relevant baselines (though some relevant similar approaches should be included). In particular, the diversity-promoting metric should be used as a baseline as it is a very well known approach. The fact that the method would be available to the community would allow additional research to be performed.

Reasons to reject

While the paper is fairly clear, there are some missing baselines that could be considered. Furthermore, some of the analysis appears to not follow the obvious inconsistencies in the metrics presented. In particular, there are both automated and human-evaluated metrics and their results do not correlate.

Questions for the Author(s)

Question A: It is interesting that your ablation studies show inconsistent behavior between BLEU and the other metrics. However, this just leads to more questions about how these automated metrics relate to the human evaluations you also did. Tables 1 and 3 also show some inconsistencies, for example BLEU and Rouge do show that your method does best, but the human evaluation of fluency does not. This suggests that BLEU (and ROUGE) do not correlate well with the human evaluation. I would expect you to comment on this and point out how to address it in future studies.

Question B:

Question C:

Missing References

There are a number of relevant papers which should be cited. Most similar is probably https://arxiv.org/abs/1510.03055 - "A Diversity-Promoting Objective Function for Neural Conversation Models". This is solving a very related problem with a somewhat similar approach, it is disappointing that it was not included as it is a very well known cite. There are a number of follow-on works from the same lab and others which should also probably be included.

Typos, Grammar, Style, and Presentation Improvements

There are many typos, among the most egregious and easily noticed (as they are in section headings):

Lines 229, 248, 263: "Weihgt" --> "Weight"

There are also a number of grammatical mistakes, for example:

Lines 116-118: "Our approach [...] effectively reduce hallucinations" (should be "reduces")

Line 199: "We will detail introduce"

Furthermore, there are a number of inconsistencies in the presentation. For example, the authors talk about how they will focus on evaluating their method in terms of the effectiveness of improving fluency and faithfulness, but the human evaluation contains a third metric, informativeness. Why not include that as well as a focus? If it is not a focus, then why include it at all? How are the automated metrics related to the human ones?

Soundness: 3
Excitement (Long paper): 3.5
Reviewer Confidence: 4
Recommendation for Best Paper Award: No
Reproducibility: 4
Ethical Concerns: No

Review #3

What is this paper about and what contributions does it make?

Authors tackle the problem of hallucinations in the context of knowledge-grounded dialogue systems. Such systems produce a response given the dialogue context and a piece of retrieved relevant knowledge. The response should be faithful to the retrieved knowledge and should not hallucinate.

Authors propose a simple but effective tweak to the training loss. The main idea is to weigh the tokens in response differently based on whether these tokens are relevant to the knowledge or not. Knowledge-aware tokens (i.e., tokens semantically similar to the knowledge) should have more contribution towards the overall loss as compared to irrelevant tokens.

Authors propose various schemes for deciding the token's weight contribution towards the loss. Authors evaluate their approach on 2 knowledge-guided dialogue datasets (Wizard of Wikipedia and FaithDialog). Evaluation is performed with automatic metrics as well as human judgments and results show improved faithfulness of response to the knowledge. Ablation studies are performed to identify which weighing schemes help the most.

Reasons to accept

Reducing hallucinations in knowledge-grounded dialogue systems is an important research direction. The paper presents a simple approach that results in improved faithfulness while preserving fluency. The paper is well organized and easy to read. Evaluation protocol is sound and presents both automatic and human judgements. Ablation studies also give a good insight into the tradeoffs between various hyperparameters.

Reasons to reject

The work presents an incremental improvement over standard T5 based seq2seq dialogue models used for knowledge-grounded dialogues. The scope of the work is rather limited to improving faithfulness given selected knowledge and knowledge selection itself is out of scope.

There are a few unexplored ablations/experiments (see below) that could have made the paper stronger.

Questions for the Author(s)

A: Figure 5 seems a bit odd. Maybe the labels are wrong on this one, but it's strange to see BERTScore improving as the training data size is reducing. Is this figure accurate? If so it'll be important to see an explanation of what's going on there.

B: What are other ways to determine knowledge-aware tokens? Why are token embedding used directly instead of BERT-encodings etc? Or even simpler overlapping tokens for content words? Were these options tried out?

C: The 'adjust weight' effectively sharpens the token distribution for knowledge aware tokens. Did you consider other ways of using the 'adjust weight' (may be directly as a multiplier per token in loss function in eq 9?

D: All the weighing schemes monotonically increase the weight as a function of relevance score, do you consider learning this function directly?

Typos, Grammar, Style, and Presentation Improvements

Please run spell checker (esp. For weight) and grammar checker.

Soundness: 4
Excitement (Long paper): 3.5
Reviewer Confidence: 4
Recommendation for Best Paper Award: No
Reproducibility: 4