View Meta-Reviews

Paper ID	8344
Paper Title	Logic-Regularized Reasoning for Interpretable Fact Verification
Track Name	Main Track

META-REVIEWER #1

META-REVIEW QUESTIONS

2. Summary of the discussion among PC members. Please note whether there is broad agreement among the reviewers regarding their overall ratings (in sentiment if not scores), and key strengths and weaknesses with respect to review criteria, and any major points of disagreement among reviewers.

No discussion seemed necessary.

3. Reasons to accept the paper. Please list the key arguments for accepting the paper.

- novel ideas
- good experimentation method
- simple and reasonable strategy
- convincing results

4. Reasons to reject the paper. Please list the key arguments for rejecting the paper.

- some techniques should be better explained
- possible reproducibility issues
- writing could be improved
- discussion with some recent related work may be missing
- experiments conducted over a single dataset

View Reviews

Paper ID
8344

Paper Title
Logic-Regularized Reasoning for Interpretable Fact Verification

Track Name
Main Track

Reviewer #1

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
This paper proposes a fact verification framework that can enhance interpretability. The interpretable method LoRen decompose a claim into several phrases and use the verifications of phrases to determine the final validity of the claim.

2. {Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Good: The paper makes non-trivial advances over the current state-of-the-art.

3. {Soundness} Is the paper technically sound?
Good: The paper appears to be technically sound, but I have not carefully checked the details.

4. {Impact} How do you rate the likely impact of the paper on the AI research community?
Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. {Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation could be improved.

6. {Evaluation} If applicable, are the main claims well supported by experiments?
Good: The experimental evaluation is adequate, and the results convincingly support the main claims.

7. {Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Fair: The shared resources are likely to be moderately useful to other AI researchers.

8. {Reproducibility} Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)
Good: key resources (e.g., proofs, code, data) are available and key details (e.g., proofs, experimental setup) are sufficiently well-described for competent researchers to confidently reproduce the main results.

9. {Ethical Considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?

Fair: The paper addresses some but not all of the applicable ethical considerations.

10. {Reasons to Accept} Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
1. The idea of reveal the interpretability of fact verification is novel and the source code that will be made publicly can be used for follow-up research.
2. The authors conduct abundant experiments to demonstrate the effectiveness of proposed framework.
3. The paper is well-written and the detailed description of the method is easy to re-implemented.

11. {Reasons to Reject} Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
Some techniques (MRC, Knowledge distillation etc.) used in the proposed method should be stated more clear, such as use a graph to present the relationships between them. And this paper used many techniques to enhance the performance, I wonder how many are truly helpful.

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.
1. In case study, the method seems detect one culprit well. How the performance of the method to detect multiple culprit?
2. Are there any cases that can not identify sub-phrases? Then how to determine the validity of such claims.

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
It is recommended that the authors explain in more detail the motivation for the use of the techniques such as MRC, knowledge distillation etc.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category
Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate to high impact on more than one area of AI, with good to excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #2

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The paper proposes a new method for interpretable fact verification based on extracted claim phrases.

Given a claim and a set of sentence treated as evidence, the method first extracts key phrases in the claim, then finds the replacement in evidence for every extracted key phrase, and finally determines the verification result (SUP/REF/NEI) of the claim according to a simple logical rule: if the claim is supported by all replacements of key phrases, the claim is supported (SUP); if the claim is refuted by at least one replacement of key phrases, the claim is refuted (REF); otherwise, there is no enough information for determining the support or refutation. Exprimental results on the FEVER 1.0 benchmark demonstrate that the interpretation mechanism (given by key phrases and their corresponding replacements in evidence) plays an important role in making the proposed method outperform state-of-the-art methods for fact verification.

2. {Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Good: The paper makes non-trivial advances over the current state-of-the-art.

3. {Soundness} Is the paper technically sound?
Fair: The paper has minor, easily fixable, technical flaws that do not impact the validity of the main results.

4. {Impact} How do you rate the likely impact of the paper on the AI research community?
Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. {Clarity} Is the paper well-organized and clearly written?
Fair: The paper is somewhat clear, but some important details are missing or unclear.

6. {Evaluation} If applicable, are the main claims well supported by experiments?
Good: The experimental evaluation is adequate, and the results convincingly support the main claims.

7. {Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Good: The shared resources are likely to be very useful to other AI researchers.

8. {Reproducibility} Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)
Poor: key details (e.g., proof sketches, experimental setup) are incomplete/unclear, or key resources (e.g., proofs, code, data) are unavailable.

9. {Ethical Considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Excellent: The paper comprehensively addresses all of the applicable ethical considerations.

10. {Reasons to Accept} Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
(1) A new method for interpretable fact verification is proposed, based on a simple but reasonable logical rule about key phrases in the claim and their corresponding phrases in evidence.
(2) Sufficent evaluation is presented to show that the proposed interpretation mechanism further improves fact verification performance beyond state-of-the-art methods for fact verification.

11. {Reasons to Reject} Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
(1) The reproducibility is problematic.
(2) The paper is rather hard to digest with some important details missing especially on culprit prediction

and probing question generation.

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.

(1) In the first paragraph of the page 3, the probability p(z_i|c,w_i,E) of the veracity of z_i conditioned mainly on a claim phrase w_i is mentioned, but this term p(z_i|c,w_i,E) is not used anywhere else. How to compute p(z_i|c,w_i,E)?

(2) The definition and computational method for culprits is crucial in estimating the reported CULPA metric. But how is culprit exactly defined and how to compute culprits in the proposed method?

(3) In the paragraph about probing question generation in $4.2, it is said that a question is generated for one extracted claim phrase, but details are missing. What is the template for constructing these questions?

(4) In the paragraph about MRC training in $4.2, it is said that the adopted generative MRC model outputs claim phrases W_c. Is it true? Why does the MRC model not output the replacements W_E for claim phrase?

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.

As far as I know, the proposed interpretation mechanism based on key phrases in claim and their corresponding replacements in evidence is novel. The proposed techniques for implementing the interpretation mechanism while outputting the final SUP/REF/NEI prediction also seem to be sound. One issue about the proposed techniques is the reproducibility. Since there are multiple models employed in the proposed method for fact verification, such as the T5-based model for generating probing questions and the generative MRC model for computing phrase replacements in evidence, the training/dev/test data for these models need to be available, otherwise the evaluation results are impossible to reproduce. The supplemental material only provides the code but not any data. The promise of releasing the code after acceptance is insufficient. The releasing of the data for probing questions and phrase replacements in evidence is key to guarantee reproducibility.

Regarding the clarity, the paper looks in good shape but is rather hard to digest, mainly due to missing some important details especially on culprit prediction and probing question generation. Please answer my questions raised above to clarify these details. Another difficulty in digesting the paper is on understanding key notions. There are several key notions used throughout the paper, including claim phrase (in W_c), phrase replacement in evidence (in W_E), local premise, and culprit. The definitions of these notions are scattered in the paper and not given rigorously. Although a general reader like me can understand almost all notions after reading the paper several times, the authors are still able to alleviate readers' labor by presenting these notions together in a more rigorous way, especially for the notion of culprit for which I have not figured out how to compute it by now.

After reading the response, I still cannot get confirm answer to what culprits is exactly defined. It only seems that culprits amount to claim phrases in W_c. If the set of culprits is only a subset of W_c. Please define it in an explicit way. Otherwise, please do not introduce such unneccsary or even misleading notions in the paper. By considering the authors' promise on making code, data and models available, I raise the score a bit.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No

more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category
Weak Accept: Technically solid, moderate to high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #3

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
This paper attempts to decompose the claim veracity into phrase-level veracity and bring the logical constraints between them. The whole framework is based on variational inference. In the detail, a QG and MRC pipeline is used to generate phrase-aware premises and the phrase-level predictions are aggregated with soft logic constraints. Experiments on FEVER show the effectiveness of the proposed method and the phrase-level veracity prediction can be as an explanation.

2. {Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Good: The paper makes non-trivial advances over the current state-of-the-art.

3. {Soundness} Is the paper technically sound?
Excellent: I am confident that the paper is technically sound, and I have carefully checked the details.

4. {Impact} How do you rate the likely impact of the paper on the AI research community?
Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. {Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation could be improved.

6. {Evaluation} If applicable, are the main claims well supported by experiments?
Fair: The experimental evaluation is weak: important baselines are missing, or the results do not adequately support the main claims.

7. {Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Good: The shared resources are likely to be very useful to other AI researchers.

8. {Reproducibility} Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)
Excellent: key resources (e.g., proofs, code, data) are available and key details (e.g., proof sketches, experimental setup) are comprehensively described for competent researchers to confidently and easily

reproduce the main results.

9. {Ethical Considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Not Applicable: The paper does not have any ethical considerations to address.

10. {Reasons to Accept} Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
1. This core idea of this paper is clear and novel for the fact verification task based on large corpora.
2. The method is reasonable.
3. The experiments are sufficient and the details are provided.

11. {Reasons to Reject} Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
1. Some new baselines in the recent three months are missing in the main table of performance comparison.
2. There are some editing issues that should be tackled.

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.
1. Probing Question Generation: How many questions are generated for one claim phrase? Is the answer 2 (one cloze question and one interrogative question)?

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
This paper attempts to decompose the claim veracity into phrase-level veracity and bring the logical constraints between them. The idea is easy to understand and the whole process is a little complex but reasonable, given that no more fine-grained annotations are provided in the FEVER dataset. Though the working process largely relies on the existing components (MRC, QG, NLI, even the variational inference framework...), it is not trivial to combine them and make them work as a whole. The ablative experiments are solid and sufficient to show the performance of each key component.

My main concern is on the main experiment, i.e., the performance comparison between the proposed LOREN and baselines:

1. I notice that some of the ACL 2021 papers are cited by the authors. But two baseline methods are not mentioned in Table 1:
(1) Exploring Listwise Evidence Reasoning with T5 for Fact Verification (which is cited);
(2) Topic-Aware Evidence Reasoning and Stance-Aware Aggregation for Fact Verification (which is not cited).
Both of them should be included.

Further, I notice a very recent preprint on arXiv, "ProoFVer: Natural Logic Theorem Proving for Fact Verification" (https://arxiv.org/abs/2108.11357), which shares a similar idea with this submission. I understand they should be regarded as contemporary works, but I suggest the authors discuss and compare in the paper. I would like to clarify this preprint does not influence my overall recommendation.

2. As a T5 model, a BART_{base} model, and a DeBERTa are used for training (QG, MRC, NLI) during the

training of LOREN, it is better to list all the basic dependencies of both the proposed LOREN and the baselines for better comparison.

Further, some editing issues should be tackled in the next version:

1) Use a comma or period dot for the equations. For example, a comma is needed at the end of Eq. (4).
2) Introduction Para 1 Line 6: Should Figure 3 be Figure 1?
3) It might be better to shorten the Latent Model subsubsection, as the variational inference is actually not that new now.
4) Local Premise Construction Para 1: "For every claim phrases..." phrases → phrase, questions →question
5) Training Details in Supplementary, Line 5: Two "beam"s. Delete one.
6) Cite the peer-reviewed papers instead of their preprints, if any. For example, HuggingFace's Transformers: State-of-the-art Natural Language Processing is published as a demo paper in EMNLP 2020.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category
Weak Accept: Technically solid, moderate to high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #4

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
This paper presents the logic-regularized reasoning model for Interpretable fact verification. Overall this idea is interesting and the paper is well-written.

2. {Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Good: The paper makes non-trivial advances over the current state-of-the-art.

3. {Soundness} Is the paper technically sound?
Good: The paper appears to be technically sound, but I have not carefully checked the details.

4. {Impact} How do you rate the likely impact of the paper on the AI research community?

Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. {Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation could be improved.

6. {Evaluation} If applicable, are the main claims well supported by experiments?
Good: The experimental evaluation is adequate, and the results convincingly support the main claims.

7. {Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Good: The shared resources are likely to be very useful to other AI researchers.

8. {Reproducibility} Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)
Good: key resources (e.g., proofs, code, data) are available and key details (e.g., proofs, experimental setup) are sufficiently well-described for competent researchers to confidently reproduce the main results.

9. {Ethical Considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Good: The paper adequately addresses most, but not all, of the applicable ethical considerations.

10. {Reasons to Accept} Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
1. This paper presents a logical-regularized reasoning model, which helps the neural fact verification models better inference via some logic rules.
2. This paper presents better performance than strong fact verification models.
3. LOREN shows its effectiveness in evaluating the MRC quality.

11. {Reasons to Reject} Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
1. The main idea of LOREN has a close relationship with model uncertainty[1]. Thus, some related work should be discussed to make the paper more complete.
2. Some related work[1,2] about MRC quality estimation should also be discussed.

[1] AnswerFact: Fact Checking in Product Question Answering. EMNLP 2020.
[2] Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. ACL 2020.

[1] Uncertain Natural Language Inference. ACL 2020.

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.
N/A

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
This paper shows some valuable studies about logic-based reasoning of fact verification. Overall this paper conducts better performance than strong fact verification baselines and shows some interesting

experimental results. But this idea has some relations with model uncertainty, which should also be discussed in this work.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category
Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate to high impact on more than one area of AI, with good to excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted

Reviewer #5

Questions

1. {Summary} Please briefly summarize the main claims/contributions of the paper in your own words. (Please do not include your evaluation of the paper here).
The paper proposes an explainable claim verification model that decomposes the problem to the phrase-level using phrases to predict overall claim veracity, and identifies "culprit" phrases that make the claim false according to input evidence.

2. {Novelty} How novel are the concepts, problems addressed, or methods introduced in the paper?
Good: The paper makes non-trivial advances over the current state-of-the-art.

3. {Soundness} Is the paper technically sound?
Excellent: I am confident that the paper is technically sound, and I have carefully checked the details.

4. {Impact} How do you rate the likely impact of the paper on the AI research community?
Good: The paper is likely to have high impact within a subfield of AI OR moderate impact across more than one subfield of AI.

5. {Clarity} Is the paper well-organized and clearly written?
Good: The paper is well organized but the presentation could be improved.

6. {Evaluation} If applicable, are the main claims well supported by experiments?
Excellent: The experimental evaluation is comprehensive and the results are compelling.

7. {Resources} If applicable, how would you rate the new resources (code, data sets) the paper contributes? (It might help to consult the paper's reproducibility checklist)
Good: The shared resources are likely to be very useful to other AI researchers.

8. {Reproducibility} Are the results (e.g., theorems, experimental results) in the paper easily reproducible? (It may help to consult the paper's reproducibility checklist.)

Good: key resources (e.g., proofs, code, data) are available and key details (e.g., proofs, experimental setup) are sufficiently well-described for competent researchers to confidently reproduce the main results.

9. {Ethical Considerations} Does the paper adequately address the applicable ethical considerations, e.g., responsible data collection and use (e.g., informed consent, privacy), possible societal harm (e.g., exacerbating injustice or discrimination due to algorithmic bias), etc.?
Excellent: The paper comprehensively addresses all of the applicable ethical considerations.

10. {Reasons to Accept} Please list the key strengths of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
- The proposed approach is interesting and unique for the very important problem of claim verification.
- The approach is explainable, which is currently a common request among users of classification models based on neural networks.
- The paper is generally easy to follow and understand.
- Thorough and meaningful experiments.

11. {Reasons to Reject} Please list the key weaknesses of the paper (explain and summarize your rationale for your evaluations with respect to questions 1-9 above).
- All experiments were conducted on a single dataset (FEVER). From working closely with this dataset, the claims were constructed using manual and artificial ways, resulting in claims with generally fixed and clear patterns. Among those methods, text changing strategies, like negation for example, where used. This might explain why the proposed method can easily detect “culprit” phrases, since these phrases were actually manually added by original dataset creators to factual statements extracted from Wikipedia to make refuted claims. Moreover, FEVER claims are strictly about a single fact, i.e., it is hard to imagine there will be multiple “culprit phrases” in the same claim about the same fact. This issue in claims of FEVER might also justify why UnifiedQA fails drastically, since it was trained on QAs of different patterns compared to those extracted from the dataset itself and used to fine-tune the MRC model. This raises some doubt regarding the performance of the proposed method (and how robust it is) over naturally-occurring claims (e.g., in debates or social media).
- Some concepts and design decisions are hastily explained hindering understanding some aspects of the approach and implementation.

12. {Questions for the Authors} Please provide questions that you would like the authors to answer during the author feedback period. Please number them.
1. How many culprit phrases were found on average per claim from the 100 refuted claims that you manually labelled?
2. Were such (or similar) approaches proposed in other text classification problems (e.g., hate speech detection, NLI, sentiment analysis, etc.)? I think it is important to compare to some of these studies in the related work section.
3. Is there any justification behind selecting the specific heuristic rules for phrase extraction?
4. Is there a justification behind selecting DeBERTa as the NLI model?

13. {Detailed Feedback for the Authors} Please provide other detailed, constructive, feedback to the authors.
- I highly suggest testing the proposed method (and comparing it to baselines) on a dataset containing naturally-occurring/colloquial claims (e.g., tweets). I raised a concern regarding the use of FEVER with the proposed method above.
- Much more details are needed regarding why your method needs “a generative machine reading

comprehension (MRC) task," under section 3.4 . Are "WE" answers generated by MRC model or input to it? Is "WE" a set of answers (hence using the "answers" plural term in the paper)? If so, how can you use it to replace phrases for a single claim "Wc"?
- There is a lot of cross-referencing within the text. I suggest grouping some sections together to make it easier to follow what was done (e.g., MRC training, question generation, etc.) mainly sections 3.4 and 4.1
- Please clarify what "Implementations" mean in section 3.4.
- "However, in our setting, it is difficult to compute the exact posterior p(z|y, x)" where is this probability in the previous equations in this section?
- I suggest clarifying what each of the three values represent/capture here "of each phrase wi … as a three-valued latent variable"
- Please indicate what Dkl in eq 4 stands for.
- I think this sentence from the intro should point to Figure 1 and not 3 "Donald Trump winning the election, as shown in Figure 3"

** After response
I thank the authors for the detailed and convincing responses to my comments. I support the acceptance of the paper. I would like to give some final suggestions:
- As stated in my original review and clarified by the authors in their answer F.A1, FEVER claims generally have a single culprit (AVG is 1.26 according to F.A1 by the authors). I think AVG. number of culprit (and this caveat/design choice of FEVER claims) should be clearly stated in the paper to provide future justification to why this system might/mightn't be effective over claims of many culprits per claim.
- I suggest for future work that the authors look at and experiment with the latest version of FEVER (FEVEROUS).
- Please clarify how the MRC component works exactly in the paper itself.
- Regarding F.A3, my main concern is the justification behind selecting these rules. I am not commenting on the effectiveness of the rules, but would like to know how the authors came-up with them, and if there is any theoretical background/practical justification for that. I suggest the authors should address this in the paper itself.

14. (OVERALL EVALUATION) Please provide your overall evaluation of the paper, carefully weighing the reasons to accept and the reasons to reject the paper. Ideally, we should have: - No more than 25% of the submitted papers in (Accept + Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 20% of the submitted papers in (Strong Accept + Very Strong Accept + Award Quality) categories; - No more than 10% of the submitted papers in (Very Strong Accept + Award Quality) categories - No more than 1% of the submitted papers in the Award Quality category
Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate to high impact on more than one area of AI, with good to excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

20. I acknowledge that I have read the author's rebuttal and made whatever changes to my review where necessary.
Agreement accepted