HITS
Heidelberg Institute for
Theoretical Studies

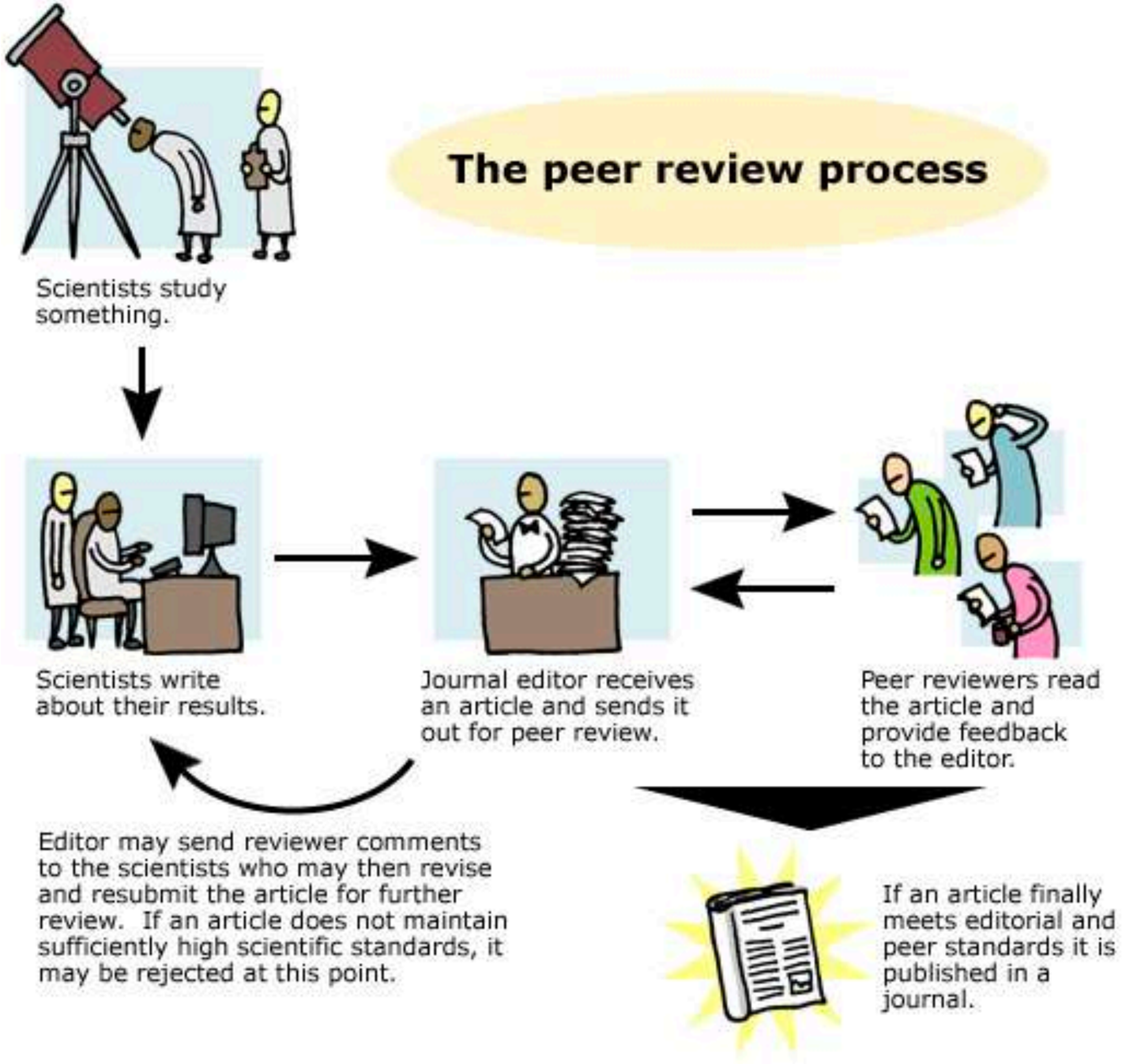
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Mini Lectures

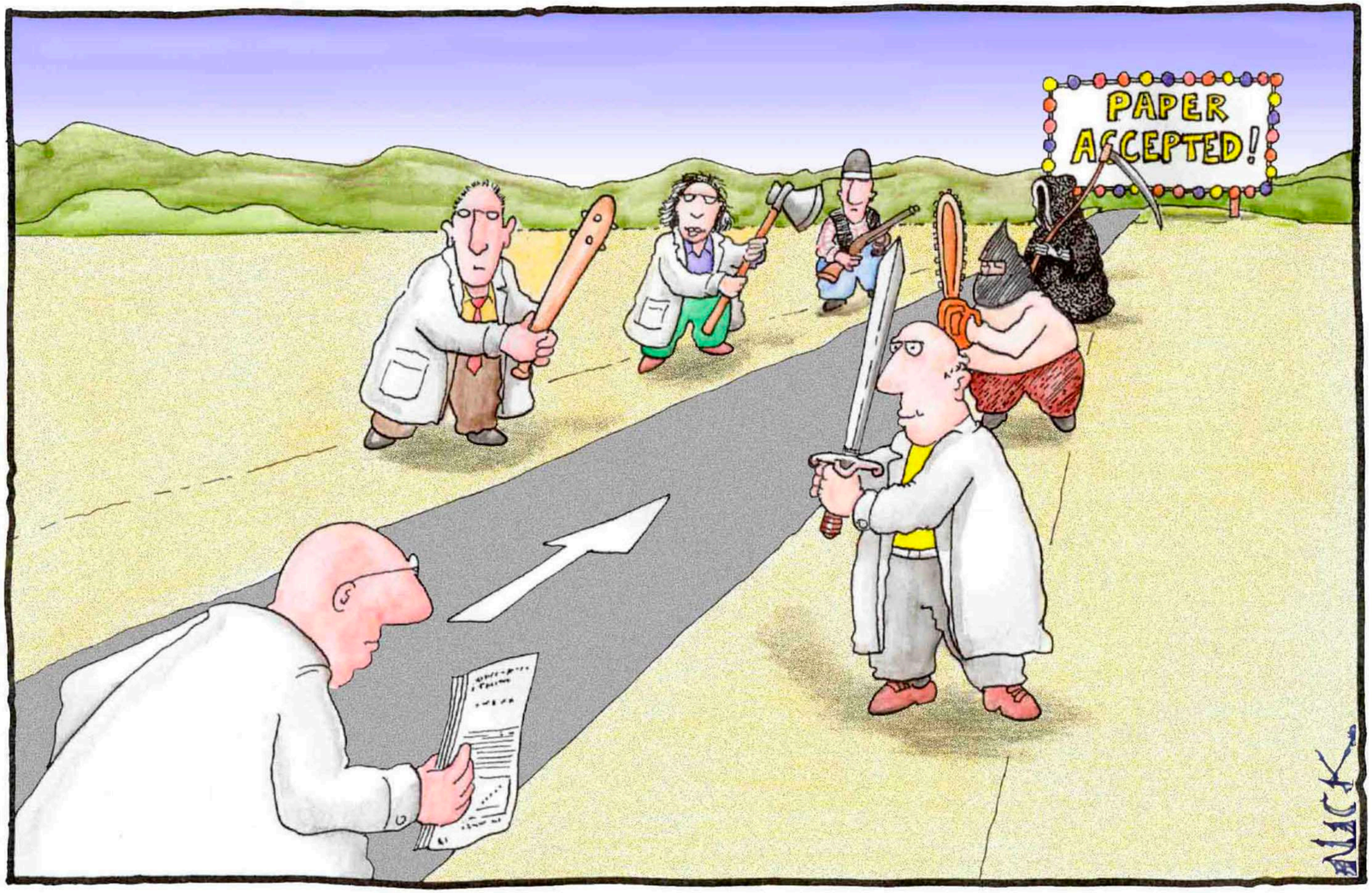
Lecture 1 - How to write a paper review

Natural Language Processing Group

Dr. Wei Zhao 31/10/2023



https://astrophysicatauricum.org/index.php/aat/peer_reviewing_process



Most scientists regarded the new streamlined peer-review process as ‘quite an improvement.’

https://s4be.cochrane.org/blog/2017/04/10/problems-with-peer-review-the-achilles-heel-of-modern-research/

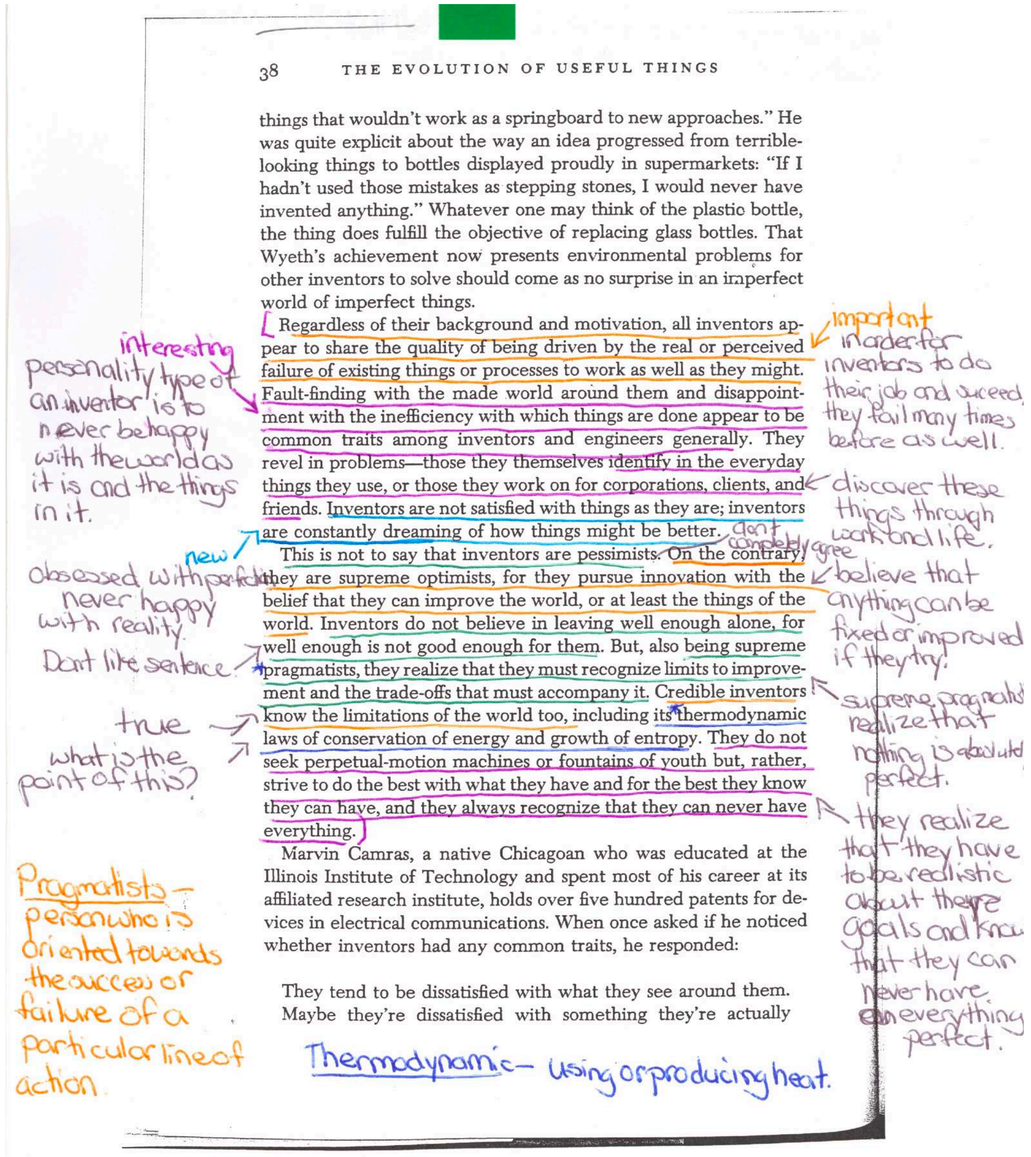
How to Become a Good Goalkeeper

- Reading is annotating
- How to structure a paper review
- Examples
- Mistakes we should avoid

Reading is Annotating

Consider the following annotations:

- Research questions
- Main results and evidence
- Strengths: novelty, good results
- Weaknesses:
  - Lack of clarity and evidence
  - Errors in methods and analyses
- Typos
- …



How to Structure a Paper Review

Introduction

The length of an introduction is usually one paragraph for a journal article review and two or three paragraphs for a longer book review. Include a few opening sentences that announce the author(s) and the title, and briefly explain the topic of the text. Present the aim of the text and summarise the main finding or key argument. Conclude the introduction with a brief statement of your evaluation of the text. This can be a positive or negative evaluation or, as is usually the case, a mixed response.

https://www.student.unsw.edu.au/structure-critical-review

How to Structure a Paper Review

Summary

1

Present a summary of the key points along with a limited number of examples. You can also briefly explain the author's purpose/intentions throughout the text and you may briefly describe how the text is organised. The summary should only make up about a third of the critical review.

https://www.student.unsw.edu.au/structure-critical-review

How to Structure a Paper Review

Critique

The critique should be a balanced discussion and evaluation of the strengths, weakness and notable features of the text. Remember to base your discussion on specific criteria. Good reviews also include other sources to support your evaluation (remember to reference).

You can also include recommendations for how the text can be improved in terms of ideas, research approach; theories or frameworks used can also be included in the critique section.

https://www.student.unsw.edu.au/structure-critical-review

What a review should include

Conclusion (Optional, often written by the Area Chairs or Journal Editors)

This is usually a very short paragraph.

- Restate your overall opinion of the text.
- Briefly present recommendations.
- If necessary, some further qualification or explanation of your judgement can be included. This can help your critique sound fair and reasonable.

https://www.student.unsw.edu.au/structure-critical-review

Examples

Introduction

Summary

Framework

Critique

Summary:
This paper introduces BARTScore, a new metric for generation based on the BART model. BARTScore frames the evaluation of generated text as a text generation problem. Additionally, the authors show prompts can be used to improve the metrics. BARTScore is evaluated on various generation tasks, including summarization, machine translation (MT) and data-to-text. The results are convincing: BARTScore and its variants outperform unsupervised metrics in nearly all cases and are comparable to the best supervised metrics on MT.

Main Review:
This paper introduces BARTScore, a new metric for generation based on the BART model. BARTScore frames the evaluation of generated text as a text generation problem. In practice, this means computing conditional probabilities under a (fine-tuned) BART model. For instance, a precision score can be obtained by computing p(h|r), the probability of generating the hypothesis given a reference text. Similarly, faithfulness can be computed as p(h|s) (generating the hypothesis given the source document) and recall can be computed as p(r|h).

BARTScore is unsupervised, in the sense that it does not require human judgements scores. However, fine-tuning the BART model on task-relevant datasets improves its correlation with human judgements. The authors experiment with tuning on CNN/DailyMail and tuning on CNN then ParaBank2. This puts BARTScore in between fully rule-based metrics (e.g: n-gram overlap such as BLEU/ROUGE), and metrics directly fine-tuned to replicate human judgements (e.g BLEURT).

The authors show that adding prompts to BARTScore can further help its correlation with human performance. This is an interesting finding in itself and in line with recent findings on the effectiveness of prompting for large LMs.

BARTScore is evaluated on various generation tasks, including summarization, machine translation (MT) and data-to-text. The results are convincing: BARTScore and its variants outperform unsupervised metrics in nearly all cases and are comparable to the best supervised metrics on MT.

Pros:

- BARTScore is highly correlated with human judgments in a variety of generation settings. It also does not require pre existing human judgements. In addition, significance testing is performed, a welcome addition that confirms the usefulness of BARTScores.
- One possible caveat is that there are many BARTScores (4 metrics * 4 models tested), and score selection can be an issue when there are no dev set human judgements available. I would like to see this mentioned, however, I do not believe this to be a dealbreaker: 1/ The metric is chosen depending on the task, not selected 2/ The BARTScore variant that performs best for each task often makes intuitive sense (fine-tuned on paraphrasing for MT, on summarization for summarization) 3/ On many tasks, several BARTScore variants are ahead of other metrics.
- The idea of using prompting to make model-based metrics more effective is interesting, and shows great results here. I believe this is a worthwhile contribution per se. It will surely lead to other papers exploring this, as the idea can be pushed further (e.g: continuous prompts, etc.).
- The paper is very well-written. The Preliminaries section in particular is great and should be required reading for researchers focused on generation metrics.
- The code for BARTScore is already published, with a clear README. There is also a leaderboard for metric evaluation. Both of those will be very valuable to the community.
- The analysis, though short, has interesting findings.

Cons:

- It would be great to see some analysis on the impact of model choice to understand the generality of BARTScore. In particular, I am curious about:
  1. How well does this work if BART is replaced with another, worse-performing model? Using older models to evaluate the quality of more recent ones can lead to issues as shown in BLEURT. I do not believe 4.4.1 fully addresses this given many evaluated models are either BART or older.
  2. Are the scores improved on some tasks due to the base metric model being the same as the models evaluated (e.g Factuality evaluation of BART based outputs)?

https://openreview.net/forum?id=5Ya8PbvpZ9

Introduction

The length of an introduction is usually one paragraph for a journal article review and two or three paragraphs for a longer book review. Include a few opening sentences that announce the author(s) and the title, and briefly explain the topic of the text. Present the aim of the text and summarise the main finding or key argument. Conclude the introduction with a brief statement of your evaluation of the text. This can be a positive or negative evaluation or, as is usually the case, a mixed response.

Summary:

This paper introduces BARTScore, a new metric for generation based on the BART model. BARTScore frames the evaluation of generated text as a text generation problem. Additionally, the authors show prompts can be used to improve the metrics. BARTScore is evaluated on various generation tasks, including summarization, machine translation (MT) and data-to-text. The results are convincing: BARTScore and its variants outperform unsupervised metrics in nearly all cases and are comparable to the best supervised metrics on MT.

2 is missing: What research questions did the paper address?

4 is missing: What is the reviewer's stance about this paper?

Summary

1 Present a summary of the key points along with a limited number of examples. You can also briefly explain the author's purpose/intentions throughout the text and you may briefly describe how the text is organised. The summary should only make up about a third of the critical review.

Main Review:

This paper introduces BARTScore, a new metric for generation based on the BART model. BARTScore frames the evaluation of generated text as a text generation problem. In practice, this means computing conditional probabilities under a (fine-tuned) BART model. For instance, a precision score can be obtained by computing p(h|r), the probability of generating the hypothesis given a reference text. Similarly, faithfulness can be computed as p(h|s) (generating the hypothesis given the source document) and recall can be computed as p(r|h). Implementation

BARTScore is unsupervised, in the sense that it does not require human judgements scores. However, fine-tuning the BART model on task-relevant datasets improves its correlation with human judgements. The authors experiment with tuning on CNN/DailyMail and tuning on CNN then ParaBank2. This puts BARTScore in between fully rule-based metrics (e.g: n-gram overlap such as BLEU/ROUGE), and metrics directly fine-tuned to replicate human judgements (e.g BLEURT). Side idea

The authors show that adding prompts to BARTScore can further help its correlation with human performance. This is an interesting finding in itself and in line with recent findings on the effectiveness of prompting for large LMs. Side idea

BARTScore is evaluated on various generation tasks, including summarization, machine translation (MT) and data-to-text. The results are convincing: BARTScore and its variants outperform unsupervised metrics in nearly all cases and are comparable to the best supervised metrics on MT. Results

Critique

The critique should be a balanced discussion and evaluation of the strengths, weakness and notable features of the text. Remember to base your discussion on specific criteria. Good reviews also include other sources to support your evaluation (remember to reference).

You can also include recommendations for how the text can be improved in terms of ideas, research approach; theories or frameworks used can also be included in the critique section.

3 & 4 are missing

2 is not a legit weakness

Pros: 1

- BARTScore is highly correlated with human judgments in a variety of generation settings. It also does not require pre existing human judgements. In addition, significance testing is performed, a welcome addition that confirms the usefulness of BARTScores.
- One possible caveat is that there are many BARTScores (4 metrics * 4 models tested), and score selection can be an issue when there are no dev set human judgements available. I would like to see this mentioned, however, I do not believe this to be a dealbreaker: 1/ The metric is chosen depending on the task, not selected 2/ The BARTScore variant that performs best for each task often makes intuitive sense (fine-tuned on paraphrasing for MT, on summarization for summarization) 3/ On many tasks, several BARTScore variants are ahead of other metrics.
- The idea of using prompting to make model-based metrics more effective is interesting, and shows great results here. I believe this is a worthwhile contribution per se. It will surely lead to other papers exploring this, as the idea can be pushed further (e.g: continuous prompts, etc.).
- The paper is very well-written. The Preliminaries section in particular is great and should be required reading for researchers focused on generation metrics.
- The code for BARTScore is already published, with a clear README. There is also a leaderboard for metric evaluation. Both of those will be very valuable to the community.
- The analysis, though short, has interesting findings.

Cons: 2 Information question

- It would be great to see some analysis on the impact of model choice to understand the generality of BARTScore. In particular, I am curious about:
  1. How well does this work if BART is replaced with another, worse-performing model? Using older models to evaluate the quality of more recent ones can lead to issues as shown in BLEURT. I do not believe 4.4.1 fully addresses this given many evaluated models are either BART or older.
  2. Are the scores improved on some tasks due to the base metric model being the same as the models evaluated (e.g Factuality evaluation of BART based outputs)?

Critique

The critique should be a balanced discussion and evaluation of the strengths, weakness and notable features of the text. Remember to base your discussion on specific criteria. Good reviews also include other sources to support your evaluation (remember to reference).

You can also include recommendations for how the text can be improved in terms of ideas, research approach; theories or frameworks used can also be included in the critique section.

Weaknesses: First, the reporting of the results and the comparisons to previous methods are unfair and potentially misleading. For some of the results, notably Table 4 for summarization evaluation, BartScore gets an advantage in being able to select different variants of itself for the different subtasks, as described by the paper. This is in comparison to other methods which must stick to one variant, such as ROUGE-1 or ROUGE-L. Second, many of the previous measures presented are not designed to measure the stated qualities. For example, ROUGE-1 was never meant to measure coherence, or indeed factuality, or fluency. Thus the table is misleading and overstates the advantage that BARTScore variants have over previous methods.

What is a legit weakness?

Critique

The critique should be a balanced discussion and evaluation of the strengths, weakness and notable features of the text. Remember to base your discussion on specific criteria. Good reviews also include other sources to support your evaluation (remember to reference).

You can also include recommendations for how the text can be improved in terms of ideas, research approach; theories or frameworks used can also be included in the critique section.

L.177: [Card, 2023] is definitely not the first one to suggest using lexical substitutes for LSCD. See, at the very least, https://aclanthology.org/2022.lchange-1.17/

L.253: I'd restrain from using adjectives like "ideally" when describing your own method.

What is a legit negative comment? It should be supported by evidence.

Critique

The critique should be a balanced discussion and evaluation of the strengths, weakness and notable features of the text. Remember to base your discussion on specific criteria. Good reviews also include other sources to support your evaluation (remember to reference).

You can also include recommendations for how the text can be improved in terms of ideas, research approach; theories or frameworks used can also be included in the critique section.

Overall, the paper provides extensive information on the speed of linguistic change which is also backed up by many examples, however, in my opinion the structure of the text could be improved. There are two key findings in the paper, and the text structure should reflect that (for example, having two sections with subsections instead of four "equal" sections). Also, there is no clear distinction between observation and explanation, e.g. in chapter 3 and 4 the observations in chapter 3 are all phenomena in English speaking countries, whereas the explanation in chapter 4 starts first by describing and explaining linguistic change in Scandinavia, then explaining changes in England. I believe it

What is a legit recommendation?

Conclusion

This is usually a very short paragraph.

- Restate your overall opinion of the text.
- Briefly present recommendations.
- If necessary, some further qualification or explanation of your judgement can be included. This can help your critique sound fair and reasonable.

Decision: Accept (Poster)

It is written by the Area Chair

Comment:

This paper proposes a framework for evaluation of text generation (BARTScore) by modeling evaluation as a text generation problem, built on top of the BART text generation algorithm. Similar to BERTScore, which is based on BERT, the metric proposed in this paper requires no training. BARTScore is also shown to improve with using textual prompts and with fine-tuning on downstream domain tasks, showing good results across the board for evaluating various generation tasks including MT and summarization, approaching evaluation metrics such as COMET which are supervised on human assessments.

All reviewers feel positively about this paper. While the underlying idea is not particularly creative, this is a solid paper which proposes a simple and effective approach for evaluation of text generation with convincing results and which is likely to be impactful.

Mistakes We Should Avoid

- Don’t confuse “must-have” with “nice-to-have”.
- Avoid harsh language. That demoralizes authors.
- Avoid writing a lukewarm review. You should take a side (positivity or negativity).

Take-aways

Things to remember

- Reviewers are goalkeepers in academia
- Learn what/where to annotate while reading papers
- Use framework to structure a paper review
- Some mistakes to avoid
- Practice, practice and practice