How to evaluate a research paper

Antoine Bosselut

École polytechnique fédérale de Lausanne

Considerations

- **Setting:** Pretend you're a conference or journal reviewer
  - Help authors improve their papers through feedback
  - Provide recommendations to conference organisers / journal editors about whether papers are worthy of acceptance in their current form
- **Goal**: Convince the AC (aka me!) about paper's quality
  - Be rigorous and opinionated
  - Back up analysis with points from the paper
- **Key Questions:**
  - What are the claims made by the authors regarding their work?
  - Do the contributions made by this work support the claims of the authors?

Contributions

- **Provide your impression of the main contributions of the work**
  - Do the authors make a theoretical contribution?
  - Do they establish a new formalism?
  - Do they curate a new datasets?
  - Do they design a new model?
  - Do they reproduce existing work?
  - Do they provide novel empirical insights into a known problem?
  - Do they perform interesting analyses of data, models, etc.
  - etc.

Every paper is different!

Considerations: Motivations

- What do the authors identify as the problem they are trying to:
  - Formalize
  - Solve
  - Analyse
  - Measure

- Possible discussion questions:
  - Does this motivation make sense?
  - Does this seems like a real problem?

Considerations: Approach

- **What do the authors create to tackle the problem they outlined as their motivation?**

- How do the authors formalise the problem into an experimental or analytical setting?
- What evaluations are developed to test their hypotheses?
- What methods do the authors devise to solve this problem?
- What datasets are created / collected?
- What other resources are conceptualised?

Considerations: Experimental Setup

- **What do we need to know to understand the results that come next?**

- What datasets do the authors use for training and evaluation (if not a central contribution)?
  - Are these reasonable data sources for this task?
  - Are the evaluation metrics suitable for the claims being tested?
- What are training hyper parameters?
- What simplifications or “shortcuts” do the authors make to go from their idealised setting to their experimental reality?
  - Smell test: do any of these “go too far”
- What baselines do they compare to?
  - Are they strong enough? Are they missing key baselines?

Considerations: Experimental Results

- **What are the main results the authors present?**
  - What are the takeaways from these results?
  - What are pitfalls in interpreting these results? Have the authors interpreted their own results correctly?
  - Are these results expected / surprising / underwhelming?
  - Again, are the evaluation metrics suitable for the claims being tested?

- **Do the authors spend enough time reconciling surprising results?**
  - Have they focused follow-up analyses on the right parts of their main findings?

Considerations: Analysis

- **What analysis did the authors run?**
  - Was it the “right” analysis?
  - Do you find the experimental design convincing?
  - Do you find the result convincing?
  - What would have made it stronger?

- **What analyses are missing?**
  - Why would this analysis strengthen the contributions of the work?
  - How difficult would it have been to conduct this analysis?

Considerations: Miscellaneous

- **Reproducibility:**
  - *Experimental*: Did the authors release code and identify datasets used?
  - *Conceptual*: Could the work be reproduced based on the information provided in the paper, appendix, and written materials?

- **Ethics:**
  - Are there potential ethics issues with this work?
  - Does the work mitigate, or at least, address these issues?
  - Ethics is broad (inequity, fairness, environmental, labor, etc.)

Evaluation: Strengths

- Which of the contributions of this work are strengths?
- What impresses you about this work?
- **Do the contributions of this work support the claims of the authors?**

- A contribution is not necessarily a strength
  - A new model can be poorly motivated or very incremental
  - A new dataset can be poorly filtered / curated
  - A new formalism can be conceptually flawed
  - An analysis can be mistaken
  - etc.

Evaluation: Weaknesses

- What parts of this work could have been improved?
- Which contributions did you find underwhelming?
- What was missing to make this work even better?
- Were the claims made by the authors overstated?

- **Note**: contributions can be weaknesses if they don't match the claims
  - A flawed experimental analysis
  - A poorly constructed problem statement
  - A model with significant methodological flaws

Bring it all together

- **Step 1: Reading**
  - Identify the claims made by the authors about their work
  - Identify the contributions of the work described by the paper
    - Use the considerations outlined in previous slides as a guide for what to focus on!
- **Step 2: Evaluating**
  - *Strengths:* Identify which contributions of the work are important
  - *Weaknesses:* Identify which parts of the work could have been better

Reviewing Resources

- https://soundcloud.com/nlp-highlights/77-on-writing-quality-peer-reviews-with-noah-a-smith
- https://acl2017.wordpress.com/2017/02/23/last-minute-reviewing-advice/
- https://sites.umiacs.umd.edu/elm/2016/02/01/mistakes-reviewers-make/
- http://luthuli.cs.uiuc.edu/~daf/CVPR21Training.html
- https://hackingsemantics.xyz/2020/reviewing-models/
- https://hackingsemantics.xyz/2020/reviewing-data/

- ACL 2020 Tutorial "Reviewing NLP" given on July 5 2020
  - https://slideslive.com/38928627/reviewing-natural-language-processing-research
  - https://github.com/reviewingNLP/ACL2020T3material

- Feel free to share your own!