• Tutorial On Reviewing for NLP notes

July 5th, 2020
Attendance:At most 55ish people (?), to 30 at the end

Poll Questions (please put an "x" next to your chosen answer)

Did you watch the videos at least partially?
Yes xxxxx
No x
Partially xx

*** Did you have a look at the slides?**
Yes xxxxx
No x
Partially xx

*** How often have you reviewed as primary reviewer?**
never xxxx
once x
twicex
five times x
more than 8 times x

*** How often have you reviewed as a secondary reviewer?**
never xxx
once xx
twice x
five times
more than 5 times xx

Which review do you think is the best one?
Review 1
Review 2 xxxx
Review 3
Review 4x

Do you think a bad paper deserves the same amount of attention and efforts as the others?
Yes x
Noxx

Do you usually check for plagiarism?
Yes
No xxxx

Do you usually check the provided data (if any)?
Yes xxxx
No

Do you usually try to re-run code (if provided)?
Yes x
No xxx

Do you think paper 42 should be accepted?
Yes
No xx x x x

Open questions (please feel free to contribute opinions and experiences...)

- Is it acceptable to stop reviewing a paper when one find a fatal flaw, e.g. ethics breach, plagiarism? Which "fatal flaw" would qualify as a deal breaker for reviewing?
plagiarism
ethics (inappropriate inference, research malpractice)

- How do you recruit secondary reviewers, and how do you collaborate with them towards the review?
area of interest, discuss the work, merits and demerits of the paper
- Should reviewer training be mandatory for major conferences?
absolutely; paper-reviewer matching, paper-bidding are critical steps in this voluntary effort
- How do you handle reviewing a paper for which you feel under-qualified?
recruit as many sub-reviewers and consult with them

Free space for additional questions and comments

ACL report: http://acl2019pcblog.fileli.unipi.it/wp-content/uploads/2019/07/ReportACL2019ReviewingSurvey.pdf
Material: https://drive.google.com/drive/folders/1YySOUHo5Ae5Efi33SF0ffP5dsIp6NEM_
Reviewing reform: https://www.aclweb.org/adminwiki/index.php?title=Short-Term_Reform_Proposals_for_ACL_Reviewing

How can reviewers be constructive?

constructiveness parameters could be politeness, feedbacks/action points

How can junior reviewers overcome impostor syndrome? How can junior researchers become reviewers?

- A good way to gain confidence might be to act as a secondary reviewer, collaborating with another reviewer before flying solo
- approach your advisor and ask him/her whether you can be a secondary reviewer.

Given the current overlap of Deep Learning approaches and Statistical NLP, how should a new reviewer be thinking about what papers are within ACL scope? Are there papers that should be flagged as more suitable for a Machine Learning conference, for instance?

(see answers to this question in the ACL T3 Rocket Chat...)

Copy Paste from Zoom Discussion (with their consent)

Ken Church:
https://www.cambridge.org/core/journals/natural-language-engineering/article/emerging-trends-reviewing-the-reviewers-again/10CDC1D71E1AEB21456CFBDA187CBCB6

Tirthankar Ghosal : There is this recent interesting paper on anatomy of harsh peer reviews: https://www.humanities.hk/news/this-paper-is-absolutely-ridiculous-ken-hyland

Gina Levow: Plagiarism is checked in many journal submissions as part of the desk reject process.
Salam Khalifa : I think plagiarism should be checked as part of the ‘desk reject’ process. Since it can be checked automatically.
Anna Rogers: @Pedro wrt your question on sota claims that are not sota - there's a discussion of this by Matt Gardner, check it out: https://twitter.com/nlpmattg/status/1220089814717886464?s=20
How about adding a check box for the authors to give feedback. If they feel offended by a review, they can check the check box. Then the PC can sample some of the reviews and see if they agree with the authors. That way, we can measure how often this happens.

Anna Rogers
@Samujjwal: the state-of-the-art result is not at all the most important thing to look for when writing a review. Here's why: https://hackingsemantics.xyz/2020/reviewing-models/

Gina Levow

Exactly - authorship checking/overlap should be automatable. And would probably be more accurate - just to flag cases that could be checked manually.

Jin Ung Lee

What is your opinion about fully public, but anonymous reviews like on openreview.net? They also have an anonymous preprint option.

Ken Church

For many years, EMNLP had a chair and a co-chair. The chair was a rising star and the co-chair was more senior. The chair's job was to rock the boat and the co-chair's job was to keep it from flipping over

Ken Church

A bigger issue than the process is what ends up being accepted at the end of the day? How are we doing on gender, geography, seniority, topics? How does this conference compare with that conference?
Do check out the minor Europe story in section 2.1 in https://www.cambridge.org/core/services/aop-cambridge-core/content/view/10CDC1D71E1AEB21456CFBDA187CBCB6/S1351324920000030a.pdf/emerging_trends_reviewing_the_reviewers_again.pdf

Ken Church

Everyone should read Tufte
I don't think we can reduce exploratory data analysis to a checklist

Karën:

https://www.edwardtufte.com/tufte/books_vdqi

Nader Akoury

Well graphing scales, can be useful, so it cannot be a binary decision whether this is acceptable.
For example, log scale can help show the tail of a distribution, especially for histograms.

Margot:

The question is not a binary decision, but it is whether the graph is acceptable or not. If it is trying to obscure something or highlight something that is not there or whether the choice is acceptable for what is being shown.

Ken Church

It all depends on error bars. There are cases where a small diff matters. If the sample size is large enough, this could matter. It could also be a good joke if the point was how to lie with statsticsi

Nader Akoury

Well, "acceptable or not" sounds like a binary decision. I'm just stating, that this seems like a reduction. The importance is clarity, not is the scale truncated.

Margot:
@Kenneth: Yes, unfortunately, it is very easy to lie with statistics. You do not even have to lie. You can just omit something to rephrase a story.
@Nader: Something can be acceptable in one case, considering all parameteres, which is not in anouther case, considering all parameters in that case.

Nader Akoury

@Margot: I agree with that statement. It doesn't sound like that's what's being stated here (especially since the checklist is a binary decision). Maybe that's something that can be explicitly mentioned.

Ken C

I think this comment could be make more encouraging. The same comment could be posed as a leading question. Is this test directional? That makes it clear that there is a weakness, but softer

Tom Ault
Medical and psychology journals have been experimenting with "pre-cleared" study where , before conducting the research, the authors submit everything except the results and conclusions (introduction, related work, methods including the data analysis to be performed), and if the reviewers agree the study has scientific validity, the journal agrees to publish the study regardless of the outcome. The authors then conduct the research, analyze the data and publish the results. Precertification addresses publication bias. The NLP community should experiment with something similar.

Ken Church

There are managers that believe that it is a mistake to micro-manage. That is, don’t solve the problem, but state the problem and the motivations and metrics. How would we know if we are making progress? The org can solve the problem after there is agreement on what we are trying to accomplish. We don’t want to confuse activities with accomplishments
It isn’t as simple as a check list

Anna Rogers

I'm actually organizing a workshop for negative results in EMNLP this year :)

Nader Akoury

The paper "What do RNN Language Models Learn about Filler–Gap Dependencies?" actually pre-registers it experiments.

Our studies were preregistered on aspredicted.org: To see the preregistrations go to aspredicted.org/X.pdf where X ∈ {md5ax, hd2df, mp9dv, uu8b5, rj2sk}.

Stefan Grunewald

I'm curious about your opinion on replication papers. There's a strong consensus that replication is very important and worthwhile (and I agree), but at the same time it seems there can also be "uninteresting" replications. How should we judge that? I suppose it has a lot to do with recency? (Replicating a model from the 90s is probably not going to be something a lot of people care about)

Ken Church

I was an area chair for a paper that reported disappointing results. There were a bunch of suggestions that our methods worked better than they do (if you don't read the fine print). The authors thought this was a problem. The reviewers thought that it was an error not to read the fine print. The fine print says that the results work on the benchmark (and nothing else). I loved the paper, but I think we rejected it

kevin

Stefan, a good replication paper makes the case that the thing that's being replicated is *worth* being replicated. Here's a sample comment from a (my) template for reviewing replication papers:
The lack of an explicit motivation for the paper is a weakness. The paper does not present an argument for why it would be significant or innovative to reproduce the specific original paper, rather than some other paper. ("Because the shared task organizers picked it" is not an argument. On the other hand, one could certainly use whatever argument was presented by the organizers, citing their paper.)

kevin

Ken, I don't think a reasonable person could disagree that it's important to work on quality assurance and/via the area chairs. MORE important? Tough for you to demonstrate, but I would love to know the argument!