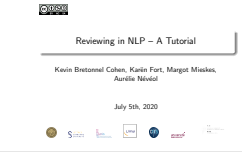
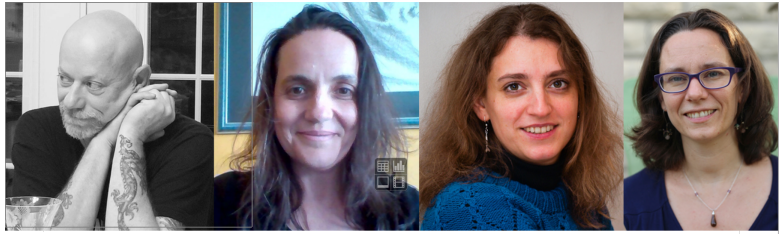
Reviewing in NLP – A Tutorial

Kevin Bretonnel Cohen, Karën Fort, Margot Mieskes, Aurélie Névéol

July 5th, 2020



Welcome



Kevin Bretonnel Cohen, UCSOM

Karën Fort, Sorbonne Université / Loria

Margot Mieskes, h_da Darmstadt

Aurélie Névéol, Université Paris Saclay, CNRS, LIMSI

Welcome

on behalf of the whole team!

Why do reviews?

Why do reviews?

take a moment to think about this

Why do reviews?

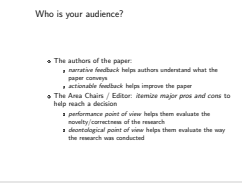
- For the good of the World
- For the good of the research community
- For your own good

Why do reviews?

- For the good of the World
- For the good of the research community
- For your own good

Who is your audience?

- The authors of the paper:
  - *narrative feedback* helps authors understand what the paper conveys
  - *actionable feedback* helps improve the paper
- The Area Chairs / Editor: *itemize major pros and cons* to help reach a decision
  - *performance point of view* helps them evaluate the novelty/correctness of the research
  - *deontological point of view* helps them evaluate the way the research was conducted



Perhaps this helps guide answers to the exercise in the next slide?

Karen: I added these, I'm not sure this is what you intended to put here

Exercise

You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Trick Question!

Exercise

You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Exercise

You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Why?

Trick Question!

Exercise

You are reviewing for a high-status conference, and for a low-status conference. How should your reviews differ?

Why?

Roles

The Gatekeeper

https://www.wired.com/2014/08/lotr-physics/

Roles

The Gatekeeper

https://www.wired.com/2014/08/lotr-physics/

Roles

The Gatekeeper

https://www.wired.com/2014/08/lotr-physics/

Not (only) YOU!

Roles

The Gatekeeper

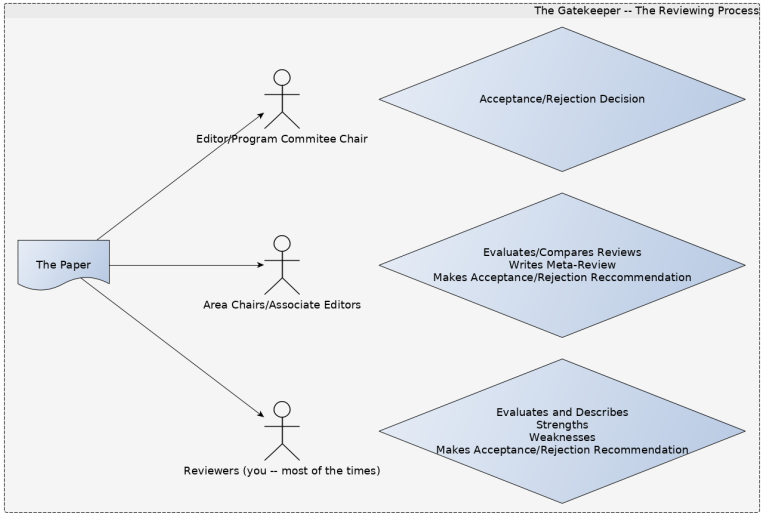
YOU SHALL NOT PASS!

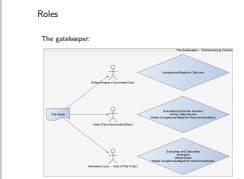
https://www.wired.com/2014/08/lotr-physics/

Not (only) YOU!

Roles

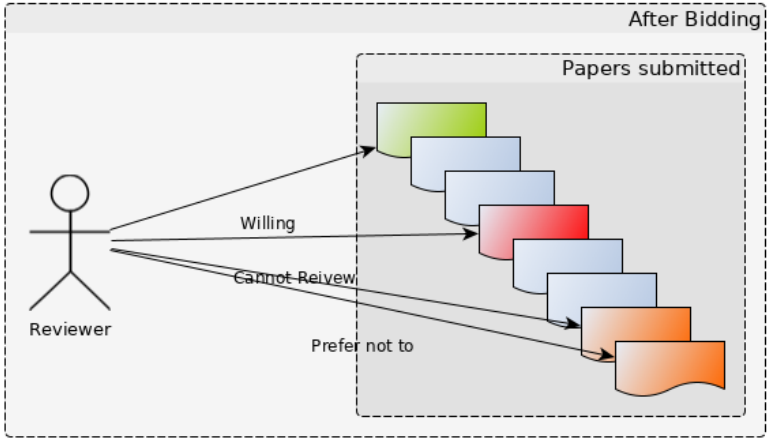
The gatekeeper:

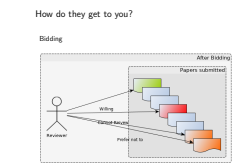




How do they get to you?

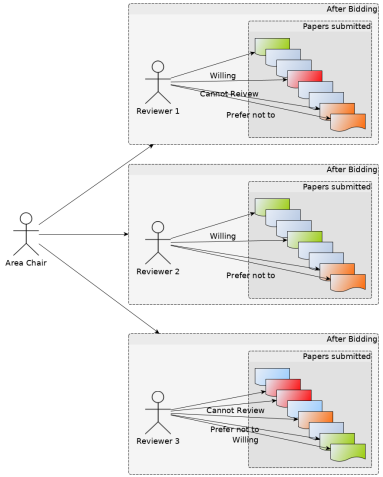
Bidding





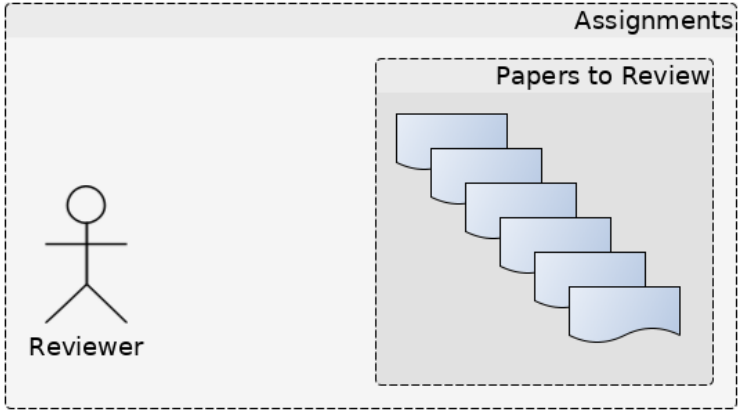
How do they get to you?

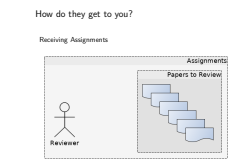
Assignment



How do they get to you?

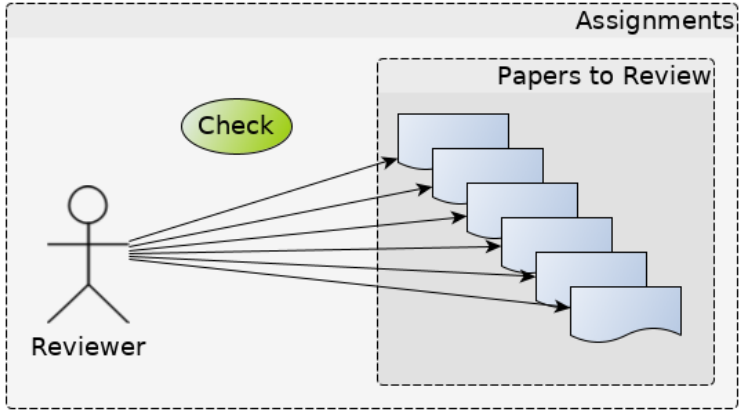
Receiving Assignments

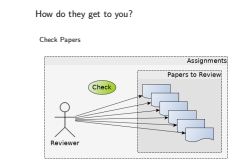




How do they get to you?

Check Papers





How do they get to you?

- Conflict of Interest?
- Concerns about your ability?
  - Topic is somehow offensive?
  - Methodology is outside of your skill scope
  - ...

Inform the Area Chair ASAP.

How do they get to you?

- Conflict of Interest?
- Concerns about your ability?
  - Topic is somehow offensive?
  - Methodology is outside of your skill scope
  - ...

Inform the Area Chair ASAP.

Karen: What about the many many cases where there are no AC (or you don't know them)?

What happens next?

Author Response Period.

https://www.pexels.com/photo/questions-answers-signage-208494/

What happens next?

Author Response Period.

https://www.pexels.com/photo/questions-answers-signage-208494/

What happens next?

Discussion Phase

https://freesvg.org/discussion

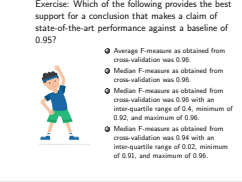
What happens next?

Discussion Phase

https://freesvg.org/discussion

Exercise: Which of the following provides the best support for a conclusion that makes a claim of state-of-the-art performance against a baseline of 0.95?

1. Average F-measure as obtained from cross-validation was 0.96.
2. Median F-measure as obtained from cross-validation was 0.96.
3. Median F-measure as obtained from cross-validation was 0.96 with an inter-quartile range of 0.4, minimum of 0.92, and maximum of 0.96.
4. Median F-measure as obtained from cross-validation was 0.94 with an inter-quartile range of 0.02, minimum of 0.91, and maximum of 0.96.



Aurélie: Should one of these options state how many experiments the average/median are computed on? Typically, 50 is better than 3...

Point: Summary statistics across multiple trials are valid only with measures of dispersion. Stable performance is preferable to highly variable performance, even with a lower measure of central tendency.

Exercise: What problem – if any – do you see with this fragment of a review?

Quantitative score: *Methodology: 1 (Methods are invalid for testing the hypothesis)*

Exercise: What problem – if any – do you see with this fragment of a review?

Quantitative score: *Methodology: 1 (Methods are invalid for testing the hypothesis)*

Exercise: What problem – if any – do you see with this fragment of a review?

Comments to authors: *The system uses rules. But, machine learning is cheaper than rules.*

Exercise: What problem – if any – do you see with this fragment of a review?

Comments to authors: The system uses rules. But, machine learning is cheaper than rules.