Gmail **Kevin B. Cohen <kevin.cohen@gmail.com>**

Your ACL 2017 Long Paper Submission (Number 473)

kanmy@comp.nus.edu.sg <kanmy@comp.nus.edu.sg> Fri, Mar 31, 2017 at 8:24 PM
To: kevin.cohen@gmail.com
Cc: kanmy@comp.nus.edu.sg

Dear Kevin Cohen:

We are sorry to inform you that the following submission was not selected by the program committee to appear at ACL 2017 (long Papers):

Two perspectives on inter-rater agreement on reproducibility in natural language processing

The selection process was very competitive. Due to time and space limitations, we could only choose a small number of the submitted papers to appear on the program. Nonetheless, We still hope you can attend the conference.

We have enclosed the reviewer comments for your perusal.

If you have any additional questions, please feel free to get in touch.

Best Regards,

Regina and Min
For the ACL 2017 scientific programme committee

==========
ACL 2017 Reviews for Submission #473
==========

Title: Two perspectives on inter-rater agreement on reproducibility in natural language processing

Authors: K. Bretonnel Cohen, Jingbo Xia, Aurelie Neveol, Prabha Yadav, Negacy Hailu, Bob Carpenter, Tiffany Callahan and Pierre Zweigenbaum

==========
REVIEWER #1
==========

Reviewer's Scores

APPROPRIATENESS: 5
CLARITY: 3
ORIGINALITY: 4
EMPIRICAL SOUNDNESS / CORRECTNESS: 2
MEANINGFUL COMPARISON: 3
SUBSTANCE: 2
IMPACT OF IDEAS OR RESULTS: 3
IMPACT OF ACCOMPANYING SOFTWARE: 1
IMPACT OF ACCOMPANYING DATASET: 2
RECOMMENDATION: 3
REVIEW DATASET: No

Comments

- Strengths:
This papers studies a problem which was not previously addressed in the literature: is it easy or not to detect if an NLP papers presents a reproducible research? This is a difficult question an this paper focuses on a prerequisite question of reproducibility: is it easy to decide if a paper gives enough information about availability of the data and/or the code needed to try to reproduced the reported experiment?

- Weaknesses:
The most important weakness about this study is the methodology used in the second experiment (on BioNLP papers). A first non-expert annotator tagged the 28 papers with an evaluation about the availability of the data and about the availability of the code. Then (lines 338-344), the first author checked the non-expert work and "When in doubt, the first author went back to the original papers". This is a common practice to use a non-expert annotation followed by an expert checking the result but it cannot be used to compute an inter-annotator agreement between the non-expert and the expert. The experiment MUST be run in the same conditions for the non-expert and the expert if an IAA is to be calculated! I think that the IAA reported lines 40-41 and lines 431-432 are useless.

Concerning the first experiment, the five annotators ran the experiment with the same settings. But, it sounds strange to ask to each annotator to work of the same data in the same order with a limited amount of time of one hour (line 248). This results in a non-homogeneous set of data: from the 9 first document are annotated five times, and the last document annotated only twice. It would have been much better to shuffle the papers for each annotator or even to ask them to spend a few more hours to annotate the full set of papers.

The analysis done of the results of this first experiment seems also over optimistic (lines 29-36): the mean Cohen's Kappa of the 10 pairs of annotators is 0.45 but the 4 best values (0.61 to 0.76) are given as an argument of substantial agreement; the same being done for observed agreement. This analysis sounds like: "If a take only the good values, my values are good!".

- Minor remarks:
* lines 103-106: I agree with the argument that in NLP, there are more data sets than in other computational areas. But I don't agree with argument (line 109-110) saying that this is also available for open source tools and code base!
* lines 233-242: very long caption with information already given in the full text: make it shorter.
* lines 256: I don't understand "usually only these two features"
* lines 260: explain the size of the first round of annotation (how many annotators?, how many papers?)
* lines 400-415: only 9 distinct points are visible on the plot.
* lines 295-299 and lines 426-429: Cohen's Kappa and agreement can be computed in a small spreadsheet (what is the benefit of using R libraries hammer here?)
* line 439: we discover that a tag "Maybe" exists, it is not described in the experiment description earlier.
* lines 450-468: how do you compute Cohen's Kappa with a non-square matrix? I guess that a third column filled with "0" is used. It should be explained.
* lines 450-468: agreement is not reported (announced line 435)

- Typos or errors:
* Repetition: line 149, "In this proposal, we focus on issues…"; line 154, "In this work, we focus on…"
* line 173: "when reproducibility is impossible" ==> "when replicability is impossible"
* line 255: "Table 4.2" no such table number in the paper

* line 384: "Table 3.1" ==> "Table 1"
* line 399: "Table 3.2" ==> "Table 4"
* line 512: "Table 4.2" ==> "Table 7"
* lines 536, 550 and 551: empty "()" bibliography reference
* line 748: typesetting problem with long URL
* lines 773 and 795: misplaced dot

==========
REVIEWER #2
==========

Reviewer's Scores

APPROPRIATENESS: 5
CLARITY: 4
ORIGINALITY: 4
EMPIRICAL SOUNDNESS / CORRECTNESS: 3
MEANINGFUL COMPARISON: 1
SUBSTANCE: 4
IMPACT OF IDEAS OR RESULTS: 4
IMPACT OF ACCOMPANYING SOFTWARE: 1
IMPACT OF ACCOMPANYING DATASET: 3
RECOMMENDATION: 3
REVIEW DATASET: No

Comments

- Strengths:

It's a good start to investigate whether it's possible to explore reproducibility of research results in nature language processing (NLP).

- Weaknesses:

1. Experiment details:
a) Chapter 3.1 describes how many annotators annotated each of the papers. In other words, each annotator didn't annotate all papers provided. Why not? If each annotator annotated all papers, there can be more pairwise agreements.

b) In computing pairwise agreements, annotators should work under the same setting. But in chapter 2.5, a non-expert annotated papers first, then an expert checked
non-expert's results. Agreement between the non-expert and the expert were calculated. non-experts and experts work in different settings.

2. The paper is well written overall, but there are several obvious written errors, which effects reviewers' impression on the paper.
a) Chapter 2.6: no contents.
b) "??" in line 435, "???" in line 549, "???? pages -." in line 689, and "()" occur.
c) Bad table and figure citations, "Table 4.2" in line255, "Figure 3.1" in line 385,
and "Table 4.2" in line 512.
d) Line 65-66: two "nonymized".

- General Discussion:

Disagreement among human raters raises concern about whether it's possible to explore reproducibility of research results in NLP. To investigate the

agreement among human raters on the reproducibility of research results, two aspects of experiments on inter-rater agreement are carried out. Results with moderate and even substantial agreements establish solid base for further research on reproducibility in NLP.

The work is pretty good overall, but there are some experiment detail errors and written errors. It can be better.

==========
REVIEWER #3
==========

Reviewer's Scores

APPROPRIATENESS: 4
CLARITY: 3
ORIGINALITY: 2
EMPIRICAL SOUNDNESS / CORRECTNESS: 2
MEANINGFUL COMPARISON: 3
SUBSTANCE: 1
IMPACT OF IDEAS OR RESULTS: 2
IMPACT OF ACCOMPANYING SOFTWARE: 1
IMPACT OF ACCOMPANYING DATASET: 2
RECOMMENDATION: 3
REVIEW DATASET: Yes

Comments

- Strengths:

The topic "reproducibility in researches" is an important topic not only in the NLP community. This is definitely worth to be studied.

- Weaknesses:

The authors presented very preliminary results with too little data to support their claims.

- General Discussion:

In their experiments, they reported that even with simple questions the IAA was 0.45 with a higher score in a subset of the annotators (4/10). I don't think that with these numbers the authors can conclude:

"p.7: we found that high inter-annotator agreement on tasks related to assessing the reproducibility of research in natural language processing can be achieved relatively quickly."

The central questions to the annotators were:

"p.3: They characterized each paper with respect to two common features found in studies of reproducibility in computational research: the availability of source code/the system, and the availability of data."

Another strong claim was: "p.7: Here, the finding was that specific domain expertise is not necessarily required". But this claim is supported only by a tiny experiment with one of the authors double-checking some annotations.

The example presented in the lines 273-279 seems to suggest that some knowledge

about the NLP domain is cleary necessary since the annotators should know about Penn Treebank resource.

The description of the experiments and the presentation of the results are somehow confusing. First, many broken cross-references (wrong number of figures and tables) and references (empty citations). Second because of the setup of the experiment and the expected results for each variation were not precisely formulated. For instance, in lines 337-344 the authors describe how one of the authors double checked data from the annotators, but how do they take the original annotation and the revised ones into account?

Lines 45-49 - I would expect to find data in a non-proprietary format.

Line 171 - broke citation.

Line 233-242 text in the caption is already discussed in the main text.

Line 384 : there is no Figure 3.1

Line 399 : there is no Table 3.2

Line 435 : broke citation

Line 536 : broke citation.

Line 549 : broke citation?

Line 550-551 : broke citations.

Line 688 : broke reference.

Line 572 : strange begining of sentence.

--
ACL 2017 - https://www.softconf.com/acl2017/papers