Reproducibility in Biomedical Natural Language Processing

Kevin Cohen, Aurélie Névéol, Jingbo Xia, Negacy Hailu, Lawrence Hunter, Pierre Zweigenbaum

Reproducibility in Biomedical Natural Language Processing

K Bretonnel Cohen, PhD[1,2], Aurélie Névéol, PhD[2], Jingbo Xia, PhD[3], Negacy Hailu, MSc[1], Lawrence Hunter, PhD[1], Pierre Zweigenbaum, PhD[2]

[1] CNRS, LIMSI, Université Paris-Saclay, F-91405 Orsay; [2]University of Colorado School of Medicine, Denver, USA; [3]Huazhong Agricultural University, Wuhan, China

Introduction

There is a growing concern that the reproducibility of much work in science in general and in the biomedical field in particular is questionable. In order to improve the robustness of methods in biomedical Natural Language Processing (NLP), we need to assess the current situation in the field, after recent experiments on a shared task suggest reproducibility cannot be taken for granted even under favorable conditions where data, code and evaluation toolkits are available[1]. Nonetheless, data and code availability is the first requirement to enable reproducibility.

Methods

Herein, we apply a protocol previously put forth to assess reproducibility in computer science[2]: all 29 articles in the proceedings of the 2016 BioNLP workshop are reviewed by two analysts (one expert in the field of biomedical NLP and one computer scientist) to extract information on code and data availability. The article passages supporting the evidence of availability are marked. However, the retrieval of the material and actual reproduction of experiments are not attempted at this stage, but would be impossible without data and code. The annotation guidelines are available at https://github.com/KevinBretonnelCohen/InterRaterAgreementReproducibility. They were created by writing an initial draft, doing a pilot annotation project, modifying the guidelines based on findings of the pilot project, and then doing the full annotation as described above. Inter analyst agreement is computed in terms of Cohen's kappa[3] in order to assess the reliability of the annotation process as well as the understanding of the task by the analysts.

Results and Conclusion

Although 48% of papers provided pointers to data and 61% provided pointers to code, only 21% reported both. Inter-analyst agreement was 0.57 (moderate) for identifying data and 0.63 (substantial) for identifying code. This suggests that reports of data availability might require more domain knowledge to be identified, e.g. a link to github might be easier to follow as evidence of code availability than the name of a dataset possibly unknown to the reader In addition, some disagreement on data availability occurred for subsets of MEDLINE where the expert considered data as not available if the specific subset and annotations were not explicitly linked. Although the kappa scores that we are reporting for the inter-analyst agreement are not high, those numbers badly underestimate the reliability of the analysis, as the calculation sets the value for expected agreement too high, resulting in an inappropriately low kappa value. The situation with respect to reproducibility in clinical natural language processing is arguably better than the situation in computer science, as Collberg et al. found that only 13% of papers in their study reported both code and data availability[2]. However, we believe reproducibility in biomedical NLP should receive increased attention, with the introduction of guidelines for reporting work in a reproducible way. While the availability of code and data do not necessarily guarantee that work can be reproduced, it is a necessary condition and as a community we should strive to facilitate access to code, data and experimental set-up as a way to address reproducibility issues.

Acknowledgements

This project was funded in part by the EU H2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207 and by the National Natural Science Foundation of China (Grant No. 61202305).

References

[1]Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proc. LOUHI 2016:78-84.

[2]Collberg C, Proebsting T, Moraila G,Sankaran A, Shi Z, Warren AM. Measuring reproducibility in computer systems research. Technical report, Department of Computer Science, University of Arizona. 2014.

[3]Cohen, J. (1968). "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit". Psychological Bulletin. 70 (4): 213–220.