Reproducibility in Computational Linguistics: Are We Willing to Share?

Martijn Wieling
University of Groningen
Center for Language and
Cognition Groningen
wieling@gmail.com

Josine Rawee
Master's student
University of Groningen
Center for Language and
Cognition Groningen
josine@rawee.nl

Gertjan van Noord
University of Groningen
Center for Language and
Cognition Groningen
g.j.m.van.noord@rug.nl

This study focuses on an essential precondition for reproducibility in computational linguistics: the willingness of authors to share relevant source code and data. Ten years after Ted Pedersen's influential "Last Words" contribution in Computational Linguistics, we investigate to what extent researchers in computational linguistics are willing and able to share their data and code. We surveyed all 395 full papers presented at the 2011 and 2016 ACL Annual Meetings, and identified whether links to data and code were provided. If working links were not provided, authors were requested to provide this information. Although data were often available, code was shared less often. When working links to code or data were not provided in the paper, authors provided the code in about one third of cases. For a selection of ten papers, we attempted to reproduce the results using the provided data and code. We were able to reproduce the results approximately for six papers. For only a single paper did we obtain the exact same results. Our findings show that even though the situation appears to have improved comparing 2016 to 2011, empiricism in computational linguistics still largely remains a matter of faith. Nevertheless, we are somewhat optimistic about the future. Ensuring reproducibility is not only important for the field as a whole, but also seems worthwhile for individual researchers: The median citation count for studies with working links to the source code is higher.

Submission received: 14 May 2018; accepted for publication: 1 July 2018.

doi:10.1162/coli_a_00330

1. Introduction

Reproducibility[1] of experimental research results has become an important topic in the scientific debate across many disciplines. There now even is a Wikipedia page on the topic entitled "Replication Crisis,"[2] with a description of some of the most worrying results and links to the relevant studies. In a survey conducted by *Nature* in 2016, more than half of over 1,500 participating scientists claim that there is a "significant reproducibility crisis."[3]

For computational linguistics, one might initially be optimistic about reproducibility, given that we mostly work with relatively "static" data sets and computer programs—rather than, for instance, with human participants or chemical substances. Yet, Pedersen (2008) points out in a very recognizable "Last Words" contribution in *Computational Linguistics* that it is often impossible to obtain the relevent data and software. Our study, ten years later, investigates whether this basic prerequisite for reproducibility is now in a better state.

Reproducing the outcome of an experiment is often difficult because there are many details that influence the outcome, and more often than not those details are not properly documented. Observations about reproducibility difficulties have been made frequently in the past. Bikel (2004), for instance, attempted to reproduce the parsing results of Collins (1999) but initially did not obtain nearly the same results. Bikel then continued to show that implementing Collins' model using only the published details caused an 11% increase in relative error over Collins' own published results.

Fokkens et al. (2013) report on two failed reproduction efforts. Their results indicate that even if data and code are available, reproduction is far from trivial, and they provide a careful analysis of why reproduction is difficult. They show that many details (including pre-processing, the experimental set-up, versioning, system output, and system variations) are important in reproducing the exact results of published research. In most cases, such details are not documented in the publication, nor elsewhere. Their results are the more striking because one of the co-authors of that study was the original author of the paper documenting the experiments that the authors set out to reproduce.

It is clear, therefore, that in computational linguistics reproducibility cannot be taken for granted either—as is also illustrated by recent initiatives, such as the IJCAI workshop on replicability and reproduciblity in NLP in 2015, the set-up of a dedicated LREC workshop series "4Real" with workshops in 2016 and 2018, and the introduction of a special section of *Language Resources and Evaluation* (Branco et al. 2017).

Our study extends the study of Mieskes (2017). She investigated how often studies published at various computational linguistics conferences provided a link to the data. She found that about 40% of the papers collected new data or changed existing data. Only in about 65% of these papers was a link to the data provided. A total of 18% of these links did not appear to work.

In our study, we focus on another essential precondition for reproduction, namely, the availability of the underlying source code. We evaluate how often data and source code are shared. We did not only follow up on links given in the papers, but we contacted authors of papers by e-mail with requests for their data and code as well. In

1 In line with Liberman (2015) and Barba (2018), we reject the unfortunate swap in the meaning of reproduction and replication by Drummond (2009). Consequently, with reproduction (or reproducibility), we denote the exact re-creation of the results reported in a publication using the same data and methods.

2 `https://en.wikipedia.org/wiki/Replication_crisis`.

3 `https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970`.

addition, we investigate to what extent we are able to reproduce results of ten studies for which we were able to obtain the relevant data and software. Our study is related to the study of Collberg, Proebsting, and Warren (2015), who investigated the frequency with which they could obtain the source code and data for publications in ACM conferences and journals, and whether the received code could be compiled. They found that only in about one third of the cases were they able to obtain and build the code without any special effort.

Importantly, we also evaluate (a rough indication of) the impact of each study via the citation counts of each study. Specifically, we assess whether there are observable differences in impact when comparing papers whose authors share their code directly (i.e., via a link in the paper) versus those that do not. Because we establish that papers that provide links to the code are typically somewhat more often cited than papers that do not, we hope to provide researchers in computational linguistics with additional motivation to make their source code available.

2. Methods

2.1 Obtaining Data and Source Code

The goal of this study is to assess the availability of the underlying data and source code of computational linguistics studies that were presented at two ACL conferences. We selected all full papers from the 2011 and 2016 ACL Annual Meetings (in Portland and Berlin), enabling us to compare the willingness and ability to share data for older (i.e., over 6 years ago at the time our study was conducted) versus more recent studies (i.e., about 1 year ago at the time our study was conducted).

Our procedure was as follows. For all 2011 and 2016 ACL full papers, we manually assessed whether data and/or software (i.e., source code) was used, modified, or created. For each paper, we subsequently registered whether links to data and/or the software were made available.[4] If data and/or source code were used and not made available, we contacted the first author of the study with a request for the data and/or source code (depending on what was missing).

Given that we wanted to obtain a realistic estimate of the number of authors who were willing to provide their data and/or source code, we constructed the e-mail text (included in the supplementary material, see Section 4) in such a way that the recipients had the impression that their specific study would be reproduced. Although this is not completely fair to the authors (since we only reproduced a small sample), simply asking them about their willingness to provide the data and/or source code without actually asking them to send the files would have resulted in overly optimistic results.[5] In addition, we explicitly indicated in the e-mail that one of the senders was a past president of the ACL. Given that the request for source code and/or data came from an established member of the ACL community, it is likely that our request was not dismissed easily. We will return to this point in the Discussion.

The first e-mail was sent on 9 September 2017 to the first author of each study for which data and/or source code was not available. If the e-mail address (extracted from the paper) no longer existed, we tried to obtain the current e-mail address via a Google

4 Data were also registered as being available if it could be obtained for a fee, such as data sets provided by the Linguistic Data Consortium.

5 This is exemplified by the fact that over a dozen authors replied to us indicating that they would provide us with the files before the deadline, but failed to do so.

search. In the very few cases where this did not work, we sent the e-mail to another author of the paper. If the author did not send the data and/or source code (nor replied that it was not possible to send the requested information), we sent a second and final e-mail on 24 October 2017. In contrast to the first e-mail, this second e-mail was sent to all authors of the paper, and the deadline for sending the information was extended to 19 November 2017.

A slightly different procedure was used for those authors who provided links in the paper to their source code and/or data that were no longer accessible. In that case we immediately contacted all authors with a request (similar to the other e-mail) to send us the updated link to the data and/or source code within two weeks. As these authors had already made this information available earlier, we only sent a single e-mail and no reminder.

Finally, for each of the 395 papers in this study, we obtained citation counts from Google Scholar on 10 March 2018.

2.2 Reproducing Results from Selected Studies

After having obtained the underlying data and/or code, we attempted to reproduce the results of a random selection of five studies from 2011 (Nakov and Ng 2011; He, Lin, and Alani 2011; Sauper, Haghighi, and Barzilay 2011; Liang, Jordan, and Klein 2011; Branavan, Silver, and Barzilay 2011) and a random[6] selection of five studies from 2016 (Coavoux and Crabbé 2016; Gao et al. 2016; Hu et al. 2016; Nicolai and Kondrak 2016; Tian, Okazaki, and Inui 2016) for which the data and source code was provided, either through links in the paper, or to us after our request.

Our approach to reproduce these results was as follows: We used the information provided in the paper and accompanying the source code to reproduce the results. If we were not able to run the source code, or if our results deviated from the results of the authors, we contacted the authors to see if they were able to help. Note that this should only be seen as a minimal reproduction effort: We limited the amount of human (not CPU) time spent on reproducing each study to a total of 8 hours. The results obtained within this time limit were compared with the original results of the aforementioned studies. The second author (a Language and Communication Technologies Erasmus Mundus Master student) conducted the replication using a regular laptop.

3. Results

3.1 Availability of Data and/or Source Code

The distribution of the links that were available and the responses of the authors we contacted is shown in Table 1. Whereas most of the data were already provided or uniquely specified in the paper (i.e., links worked in 64% and 79% of cases for 2011 and 2016, respectively), this was not the same for the source code (provided in 19% and 36% of cases, respectively). After having contacted the authors, and including that data and source code as well (i.e., providing the updated link or sending the data and/or source code), these percentages increased to 76% and 86% for the data availability, and 33% and 59% for the source code availability. When contacting the authors, the

6 The study of Nicolai and Kondrak (2016) was included as the authors explicitly asked if we could include them in the experimentation process.

Table 1
Distribution of data and code availability in both 2011 and 2016.

	2011: data		2016: data		2011: code		2016: code	
Data / code available	**116**	**75.8%**	**196**	**86.3%**	**48**	**33.1%**	**131**	**59.3%**
- working link in paper	98	64.1%	179	78.9%	27	18.6%	80	36.2%
- link sent	11	7.2%	15	6.6%	17	11.7%	50	22.6%
- repaired link sent	7	4.6%	2	0.9%	4	2.8%	1	0.5%
Data / code unavailable	**37**	**24.2%**	**31**	**13.7%**	**97**	**66.9%**	**90**	**40.7%**
- sharing impossible	19	12.4%	14	6.2%	46	31.7%	42	19.0%
- no reply	17	11.1%	12	5.3%	43	29.7%	32	14.5%
- good intentions	0	0.0%	2	0.9%	5	3.4%	12	5.4%
- link down	1	0.7%	3	1.3%	3	2.0%	4	1.8%
Total	***153***	***100%***	***227***	***100%***	***145***	***100%***	***221***	***100%***
No data/code used	11		4		19		10	
Total nr. of papers	164		231		164		231	

most frequent response type was that sharing was impossible due to (for example,) having moved to another institute or company and not having access to the data, being prohibited from sharing source code that used proprietory company tools, or having lost the data or source code. The second-most frequent type we observed was the absence of action. In those cases, we did not receive any reply to our e-mails. The third-most frequent response type was authors with good intentions, who replied that they were going to send the requested data and/or code, but did not end up doing so. In only a very few cases (1–2%), the link to the source code and/or data was not provided anew, if they were initially present in the paper and no longer working. The total percentage of available data and/or source code is informative, but another important measure is how often the source code and/or data were provided when it had to be requested (i.e., the sum of the sent and repaired link sent frequencies in the appropriate column in Table 1 as a proportion of the sum of these two frequencies and the number of papers in the corresponding column for which data or code was unavailable). Unfortunately, these percentages are rather low, with 32.7% for requested 2011 data, 35.4% for requested 2016 data, 17.8% for requested 2011 source code, and 36.2% for requested 2016 source code. In sum, if data and/or source code were not referenced through a link to a repository in the paper, authors will most likely not (be able to) supply this information.

Nevertheless, there is a clear improvement between 2011 and 2016. The number of papers containing a working link to source code almost doubled. Of course, the improvement can be explained at least partly by observing that it is much easier to share recent data and source code, rather than older data and code from 5 years ago.

Subsequently, another important question is, if we get access to the data and/or code, how likely is it that the results reported therein are reproducible? The following subsection attempts to provide a tentative answer to this question.

3.2 Reproducibility of Selected Studies

For the 2011 papers we selected, we were only able to reproduce the results of a single study (Liang, Jordan, and Klein 2011) perfectly (time invested: 4 hours). For the study of He, Lin, and Alani (2011), we were able to reproduce the results almost (but not

quite) perfectly: Their reported performance was 94.98%, whereas the performance we obtained was 94.8% (using the same version of the underlying MAchine Learning for LanguagE Toolkit the authors used). Interestingly, the performance was reduced to 83.3% when the most recent version of the MALLET toolkit was used (which was also noted by the authors). We were not able to reproduce any results for the three remaining 2011 studies we selected (Branavan, Silver, and Barzilay 2011; Nakov and Ng 2011; Sauper, Haghighi, and Barzilay 2011).

The results are better for the 2016 papers we selected. For the paper of Coavoux and Crabbé (2016), we were able to reproduce (within 5.5 hours) most of the results exactly as reported. There was only a single value out of ten that we were not able to compute. For the study of Gao et al. (2016), the reproduction results (obtained within 2 hours) were also similar. On the basis of all data, two accuracy scores out of four were identical, and the other two deviated by 0.7 and 1.1 points. The results for the corresponding baseline deviated by 0.3 and 1.0 points. The results regarding individual verbs (i.e., subsets of all data) were much more variable, with deviations of up to 16.7 points. However, this was caused by the smaller sample sizes (ranging from 6 to 58). For the study of Hu et al. (2016) we obtained results (within 3.5 hours) that were (almost) identical to those reported in the paper (i.e., 88.8 and 89.1 versus 88.8 and 89.3 reported in the paper). The models used by Nicolai and Kondrak (2016) took a long time to train and for this reason we were only able to validate two accuracy values (out of nine). Both values we obtained (taking 8 hours to compute) were similar, but not identical to those reported in the paper (reported performance: 98.5 and 82.3, our performance: 94.8 and 80.8). Finally, we were able to reproduce (within 3.5 hours) most of the results reported by Tian, Okazaki, and Inui (2016). Four out of six performance values were reproduced exactly, the remaining two performance values we checked only differed slightly (0.41 and 81.2% compared to the reproduced values of 0.42 and 81.1%, respectively).

In sum, we were only able to reproduce the *identical* results of a single study (Liang, Jordan, and Klein 2011). Of course some variability may be expected, due to (for example) randomness in the procedure. If we are a bit more flexible and ignore the single value we were not able to compute during the reproduction of Coavoux and Crabbé (2016) and the small sample results of Gao et al. (2016), and also ignore deviations for reproduced results of up to 2 percentage points, then two 2011 studies (Liang, Jordan, and Klein 2011; He, Lin, and Alani 2011) and four 2016 studies (Coavoux and Crabbé 2016; Gao et al. 2016; Hu et al. 2016; Tian, Okazaki, and Inui 2016) were reproduced successfully.

3.3 Citation Analysis

To see if there is a tangible benefit for authors to share the source code underlying their paper, we contrasted the number of citations for the papers that provided the code through a link in the paper to those that did not. Comparing the citation counts for the papers from 2011 showed a non-significant ($p > 0.05$) higher mean citation count for the studies that did not provide the source code compared with those that did provide the source code: $t(117.74) = -0.78$, $p = 0.44, m_{sc} = 71, m_{no\text{-}sc} = 84$. Note that the higher mean for the studies that did not provide the link to the code is caused by 12 highly cited papers. Excluding these outliers (and the single outlier from the 2011 papers that did provide a link to the code) yields the opposite pattern, with a significant *higher* mean citation count for the 2011 papers providing the source code than those that did not: $t(52.19) = 2.13$, $p = 0.04, m_{sc} = 62, m_{no\text{-}sc} = 44$. For 2016, we observe a significant difference, with a higher citation count for the papers providing the source

code than those that did not: t(115.12) = 2.1, p = 0.04, $m_{sc} = 27$, $m_{no\text{-}sc} = 15$. Excluding the outliers (9 papers providing the source code, 15 papers that did not provide the source code) strengthened this effect: t(94.68 = 3.7, p < 0.001, $m_{sc} = 14.3$, $m_{no\text{-}sc} = 7.3$). Papers providing the source code had a mean citation count almost double that of the papers that did not provide the source code.

Even though the t-test is highly robust to deviations from normality (Zar 1999, pages 127–129), we also analyzed the results using (quasi-)Poisson regression. This supplementary analysis supported the findings resulting from the t-test: when analyzing all data including outliers, the difference between the 2011 papers providing a link to the underlying source code versus those which did not was not significant (p = 0.58). For the 2016 papers, the difference was significant (p = 0.01). When excluding the outliers, the differences were significant for both 2011 and 2016 (all p's < 0.04).

Given that citation counts are highly skewed, we also compared the medians (that are influenced less by outliers). For 2011, the median citation count for the papers that provided a link to the source code was 60, whereas it was only 30 for those that did not provide a link to the source code underlying the paper. Despite the large difference in medians, this difference was not significant (Mann-Whitney U test: p = 0.15). For the papers published in 2016, the median citation count for the papers providing a link to the source code was 8, whereas it was 6 for those which did not. As with the t-test, this difference was significant (Mann-Whitney U test: p = 0.005). When excluding the outliers, the median group differences were significant for both years (all p's < 0.02).

In sum, papers that provided a link to the source code were more often cited than those that did not. Although this may suggest that providing a link to the source code results in a greater uptake of the paper, this relationship is not necessarily causal. Even though providing the source code may make it easier for other authors to build upon the approach of the other authors, it is also possible that authors who provide links to the source code may have spent more time carefully planning and working on the paper, thereby increasing the quality of the work and thus the uptake by the community.

4. Discussion

In this article we have assessed how often data and/or source code is provided in order to enable a reproducibility study. Although data are often available, source code is made available less often. Fortunately, there is a clear improvement from 2011 to 2016, with the percentage of papers providing a (working) link to the source code approximately doubling (from 18.6% to 36.2%). Unfortunately, requesting the source code (if it was not already provided) is unlikely to be successful, as only about a third of the requests was (or could be) granted. It is likely that the (relatively low) success of our requests is an upper bound. The reason for this is that we signed our e-mails requesting the data and/or source code with the name of an established member of the ACL community (a past ACL president).

Finally, even if the source code and data are available, there is no guarantee that the results are reproducible. On the basis of five studies selected from 2011 and five studies from 2016, we found that at most 60% of the studies were reproducible when not enforcing an exact reproduction. If an exact reproduction was required, then only a single study (from 2011) was reproducible. Approaches such as providing a virtual (e.g., Docker) image with all software, source code, and data, or using CodaLab worksheets as done by Liang, Jordan, and Klein (2011) might prove to be worthwhile in order to ensure a more effortless reproduction.

We would like to end with the following recommendation of Pedersen (2008, page 470) made ten years ago, but remaining relevant today:

> [Another course of action] is to accept (and in fact insist) that highly detailed empirical studies must be reproducible to be credible, and that it is unreasonable to expect that reproducibility be possible based on the description provided in a publication. Thus, releasing software that makes it easy to reproduce and modify experiments should be an essential part of the publication process, to the point where we might one day only accept for publication articles that are accompanied by working software that allows for immediate and reliable reproduction of results.

Because we established that papers that provide links to the code are typically more often cited than papers that do not, we hope to provide researchers in computational linguistics with additional motivation to make their source code available in future publications.

Acknowledgments

We thank all authors who took the effort to respond to our request for their data and code.

The (anonymized) count data, the reproduced values for the ten studies, listings of the reasons for replication failure, the R code used for the statistical analyses (including box plots), and the e-mail text requesting the data and code can be downloaded at `http://www.let.rug.nl/wieling/CL-repro/repro.xlsx`.

References

Barba, Lorena A. 2018. Terminologies for reproducible research. *arXiv preprint arXiv:1802.03311*.

Bikel, Daniel M. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511.

Branavan, S. R. K., David Silver, and Regina Barzilay. 2011. Learning to win by reading manuals in a Monte-Carlo framework. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 268–277, Portland, OR.

Branco, António, Kevin Bretonnel Cohen, Piek Vossen, Nancy Ide, and Nicoletta Calzolari. 2017. Replicability and reproducibility of research results for human language technology: Introducing an LRE special section. *Language Resources and Evaluation*, 51(1):1–5.

Coavoux, Maximin and Benoit Crabbé. 2016. Neural greedy constituent parsing with dynamic oracles. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 172–182, Berlin.

Collberg, Christian, Todd Proebsting, and Alex M. Warren. 2015. Repeatability and benefaction in computer systems research. Technical report. `http://reproducibility.cs.arizona.edu/`.

Collins, Michael. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Drummond, Chris. 2009. Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning IV*, Montreal.

Fokkens, Antske, Marieke Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the ACL*, volume 1, pages 1691–1701, Sofia.

Gao, Qiaozi, Malcolm Doering, Shaohua Yang, and Joyce Chai. 2016. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1814–1824, Berlin.

He, Yulan, Chenghua Lin, and Harith Alani. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 123–131, Portland, OR.

Hu, Zhiting, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. 2016. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 2410–2420, Berlin.

Liang, Percy, Michael I. Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 590–599, Portland, OR.

Liberman, Mark. 2015. Replicability vs. reproducibility—or is it the other way around? Language Log, 31 October; http://languagelog.ldc.upenn.edu/nll/?p=21956.

Mieskes, Margot. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia.

Nakov, Preslav and Hwee Tou Ng. 2011. Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 1298–1307, Portland, OR.

Nicolai, Garrett and Grzegorz Kondrak. 2016. Leveraging inflection tables for stemming and lemmatization. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1138–1147, Berlin.

Pedersen, Ted. 2008. Empiricism is not a matter of faith. *Computational Linguistics*, 34(3):465–470.

Sauper, Christina, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the ACL*, pages 350–358, Portland, OR.

Tian, Ran, Naoaki Okazaki, and Kentaro Inui. 2016. Learning semantically and additively compositional distributional representations. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1277–1287, Berlin.

Zar, Jerrold H. 1999. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ.