A Quantitative Study of Data in the NLP community

Margot Mieskes
Information Science
Darmstadt University of Applied Sciences
margot.mieskes@h-da.de

Abstract

We present results on a quantitative analysis of publications in the NLP domain on collecting, publishing and availability of research data. We find that a wide range of publications rely on data crawled from the web, but few give details on how potentially sensitive data was treated. Additionally, we find that while links to repositories of data are given, they often do not work even a short time after publication. We put together several suggestions on how to improve this situation based on publications from the NLP domain, but also other research areas.

1 Introduction

The Natural Language Processing (NLP) community makes extensive use of resources available on the internet. And as research in NLP attracts more attention by the general public, we have to make sure, our results are solid and reliable, similar to medicine and pharmacy. In the case of medicine, the general public is often too optimistic. In NLP this over-optimism can have a negative impact, such as in articles on automatic speech recognition[1] or personality profiling[2]. Few point out, that the algorithms are not perfect and do not solve all the problems, as on terrorism prevention[3] or sentiment analysis[4].

Therefore, important questions are, what happens to the data and how reliable are results obtained through them.

We present a quantitative analysis of how often data is being collected, how data is published, and what data types are being collected. Taken together it gives insight into issues arising from collecting data and from distributing it via channels, that do not allow for reproducing results, even after a comparably short period of time. Based on this, we open a discussion about best practices on data collection, storage and distribution in order to ensure high-quality research, that is solid and reproducable. But also to make sure, users of, i.e., social media channels are treated according to general standards concerning sensitive data.

2 Related Work

In the following we give a broad overview on reusability of published code and data sets, but also on results of actual reproducibility studies and privacy issues from various domains.

General Guidelines "One goal of scientific publication is to share results in enough detail to allow other research teams to reproduce them and to build on them" (Iorns, 2012). But even in medical or pharmaceutical research failure to replicate results can be as high as 89% (Iorns, 2012). Journals such as Nature[5] and PLOS[6] require their authors to make relevant code available to editors and reviewers. If code cannot be shared, the editor can decline a paper from publication.[5] Additionally, they list a range of repositories that are "recognized and trusted within their respective communities" and meet accepted criteria as "trustwor-

[1] https://theintercept.com/2015/05/05/nsa-speech-recognition-snowden-searchable-text/

[2] http://www.digitaltonto.com/2013/the-dark-side-of-technology/

[3] http://www.telegraph.co.uk/news/uknews/terrorism-in-the-uk/11431757/Algorithms-and-computers-wont-stop-terrorism.html

[4] http://www.nytimes.com/2009/08/24/technology/internet/24emotion.html?_r=1

[5] http://www.nature.com/authors/policies/availability.html

[6] http://journals.plos.org/plosone/s/data-availability

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 1–7,
Valencia, Spain, April 4th, 2017. © 2017 Association for Computational Linguistics

thy digital repositories" for storing data[6]. This enables authors to follow best practices in their fields for the preparation, recording and storing of data.

Study on re-usability of Code Collberg et al. (2015) did an extensive study into the release and usability of code in the domain of computer science. The authors categorized published code into three categories: Projects that were obtained and built in less than 30 minutes, projects that were successfully built in more than 30 minutes and projects where the authors had to rely on the statement of the author of the published code.

Additionally, they carried out a user study, to look into reasons why code was not shared. Reasons were (among others), that the code will be available soon, that the programmer left or that the authors do not intend to release the code at all.

Their study also presents reasons why code or support is unavailable. They found that problems in building code were (among others) based on "files missing from the distribution" and "incomplete documentation". The authors also list lessons learned from their experiment, formulated as advice to the community such as: plan to release the code, plan for students to leave, create project websites and plan for longevity.

Finally, the authors present a list of suggestions to improve sharing of research artifacts, among others on how to give details about the sharing in the publications, beyond using public repositories and coding conventions.

Re-using Data Some of the findings by Collberg et al. (2015) apply to data as well. Data has to be "independently understandable", which means, that it is not necessary to consult the original provider (Peer et al., 2014). A researcher has the responsibility to publish data, code and relevant material (Hovy and Spruit, 2016). Additionally, Peer (2014) argued, that a data review process as carried out by data archives such as ICSPR[7] or ISPS[8] is feasible.

Milšutka et al. (2016) propose to store URLs as persistent identifiers to allow for future references and support long-term availability.

Francopoulo et al. (2016) looked at NLP publications and NLP resources and carried out a quantitative study into resource re-usage. The authors suggest a resource innovation impact factor to encourage the publication of data and resources.

Gratta et al. (2016) studied the types of resources published during the previous three LREC conferences. They found that more than half (58%) of the resources were corpora. They visualized collaborations between researchers on specific resources and pointed out issues concerning the meta-data provided by data publishers.

Replication Studies in NLP Experiments in reproducing results in the NLP domain such as (Fokkens et al., 2013) are still quite rare. One reason might be, that when undertaking such projects, "sometimes conflicting results are obtained by repeating a study" (Jones, 2009). Fokkens et al. (2013) found, that their experiments were difficult to carry out and to obtain meaningful results. The `4Real` workshop focused on the "the topic of the reproducibility of research results and the citation of resources, and its impact on research integrity"[9]. Their call for papers[9] asked for submissions of "actual replication exercises of previous published results" (see also (Branco et al., 2016)). Results from this workshop found that reproducing experiments can give additional insights, and can therefore be beneficial for the researchers as well as for the community (Cohen et al., 2016).

Data Privacy and Ethics Another important aspect is data privacy. An overview on how to deal with data taken from, for example, social media channels can be found in (Diesner and Chin, 2016). The authors raise various issues regarding the usage of data crawled from the web. As data obtained through these channels is, strictly speaking, restricted in terms of redistribution, reproducibility is a problem.

Wu et al. (2016) present work on developing and implementing principles for creating resources based on patient data in the medical domain and working with this data.

Bleicken et al. (2016) report efforts on anonymization of video data from sign language. The authors developed a semi-automatic procedure to black relevant parts of the video, where named entities are mentioned.

Fort and Couillault (2016) report on a survey on the awareness and care NLP researchers show towards ethical issues. The authors scope also considered working conditions for crowd workers.

[7] `http://www.icpsr.umich.edu/icpsrweb/index.jsp`

[8] `http://isps.yale.edu/research/data`

[9] `http://4real.di.fc.ul.pt/`

Their results indicate that the majority (84%) consider licensing and distribution of language data during their work. Over three-quarters of the participants (77%) think that “ethics should be part of the subjects for the call for papers”.

3 Research Questions

In the course of this work, we looked at various aspects of experimental work:

Collection NLP researchers collect data, often without informing the persons or entities who produced this data. These data sets are analyzed, conclusions are drawn about how people write, behave, etc. and others make use of these findings in other contexts. This gave raise to the questions:

- Has data been collected?
- If the data contains potentially sensitive data, which post-processing steps have been taken (i.e. anonymization)?
- Was the resulting data published?
- Is there enough information/is it possible to obtain the data?

Replicability/Reproducibility Often data on which these studies are based, is not published or not available anymore. This can be due to various reasons[10]. Among those are, that webpages or e-mail addresses are no longer functional after a researcher left a specific research institute, after a webpage re-design some data has not been moved to the new page, and copyright or data privacy issues could not be resolved.

This gives rise to issues, such as reproducibility of research results. Original results from these studies are published and later referred to, but they cannot be verified on the original data. In some cases, data is being re-used and extended. But often only specific parts of the original data is used. Details on how to reproduce the changed data set (e.g. code/scripts used to obtain the subset) are not published and descriptions about the procedure are insufficient. This is extends the questions:

- Was previously published data used in a different way and/or extended?

These questions target at how easy it would be to follow-up by reproducing published results and extending the work. Our results give an indication on the availability of research data.

Specific to data taken for example from social media channels is another, additional aspect:

Personal Data Researchers present and publish their data and results of their research on conferences and workshops, often using examples taken from the actual data. And of course, they aim to look for examples that are entertaining, especially during a presentation. But we also observed that names are being used. Not just fairly common names, but real names or aliases used on social media. Which renders this person identifiable as defined by the data protection act below.

Therefore, we added the questions:

- Did the data contain sensitive data?
- Was the data anonymized?

These questions look at how researchers deal with potentially sensitive data. The results indicate how serious they take their responsibility towards their research subjects, which are either voluntarily or involuntarily taking part in a study.

What constitutes sensitive data? Related to the above presented questions, we had to define what sensitive data is. In a leaflet from the MIT Information Services and Technology sensitive data includes information about “ethnicity, race, political or religious views, memberships, physical or mental health, personal life (...) information on a person as consumer, client, employee, patient, student”. It also includes contact information, ID, birth date, parents names, etc. (Services and Technology, 2009). The UK data protecton act contains a similar list.[11] The European Commission (Schaar, 2007) formulates personal and therefore sensitive data as “any information relating to an identified or identifiable natural person”. And even anonymizing data does not solve all issues here, as “(...) information may be presented as aggregated data, the original sample is not sufficiently large and other pieces of information may enable the indentification of individuals”.

Based on these definitions, we counted towards the *sensitive data* aspect everything that users themselves report (“user generated web content” (Diesner and Chin, 2016)), but also what is being reported about them, e.g. data gathered from

[10] This is based on personal experience and therefore not quantified.

[11] `https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/`

Venue	# papers	# data published	Ratio
NAACL	182	57	31.3%
ACL	231	63	27.3%
EMLNP	264	81	30.7%
Coling	337	89	26.4%
LREC	744	414	55.6%
total	1758	704	40.0%

Table 1: Results of papers reporting the usage and the publication of data.

equiment such as mobile phones which allows to identify a specific person.

4 Quantitative Analysis

Our quantitative analysis was carried out on publications from NAACL (Knight et al., 2016), ACL (Erk and Smith, 2016), EMNLP (Su et al., 2016), LREC (Calzolari et al., 2016) and Coling (Matsumoto and Prasad, 2016) from 2016. This resulted in a data set of 1758 publications, which includes long papers for ACL, long and short papers for NAACL, technical papers for Coling and full proceedings for EMNLP and LREC, but no workshop publications.

Procedure All publications were manually checked by the author. Creating an automatic method proved to be infeasible, as the descriptions on whether or not data was collected, whether it is provided to the research community, through which channel etc. is too heterogeneous across the publications. We checked the abstracts for pointers on the specific work and looked at the respective sections on procedure, data collection and looked for mentions of publication plans, link or availablility of the data. This information was collected and stored in a table for later evaluation. This analysis could have been extended by contacting the data set authors and looking at the content of the data sets. While this definitely would be a worthwhile study, this would have gone beyond the scope of the current paper, as it would have meant to contact at least over 700 authors individually. Additionally, this project was intended to raise the awareness on how data is being collected and published.

Reproducibility of Results Of the 1758 publications 704 reported to have collected or extended/changed existing data[12] (approx. 40%).

Table 1 shows the results with respect to the number of publications and the number of papers reporting data usage and/or extension. LREC saw the highest number of published papers containing collected and/or published data.

Table 2 gives details about the availability of the data sets used. 468 of the 704 publications (58%) report a link where the data can be downloaded. Another 35% report no link at all and below 1% mention that the data is proprietary and cannot be published. Out of the links given, 18% do not work at all. This includes cases where the mentioned page did not exist (anymore) or where it is inaccessible. Most cases where links did not work (15.7%) were due to incomplete or not working links to personal webpages at the respective research institutions. Therefore, we looked in more detail at the hosting methods for publishing data. We found that only about 20.7% were published on public hosting services such as `github`[13] or `bitbucket`[14]. While these services are targeted towards code and might not be appropriate for data collections, they are at least independent of personal or research institute webpages. LREC publications also mention hosting services such as metashare[15], the LRE Map[16] or that data will be provided through LDC[17] or ELRA[18] (8.9%).

Category	Percentage
Link available	65.2%
Link does not work	15.7%
No Link	31.4%
On Request	1.8%
Proprietary data	< 1%

Table 2: Detailed numbers on available and working links

Responsibility towards Research Subjects Out of 704 publications about 32.8% collected or used data from social media or otherwise sensitive data as outlined in Section 3 above. Only about 3.5% of these report the anonymization of the data. In some cases it was obvious that no anonymization has been carried out, as the discussion of the data and results mentions user names or aliases, which makes the person identifiable. The remaining publications do not mention how

[12] Publications used more than one data set, therefore, sums can be more than 100%.

[13] https://github.com/

[14] https://bitbucket.org/product

[15] http://www.meta-share.eu/

[16] http://www.resourcebook.eu/

[17] https://www.ldc.upenn.edu/

[18] http://catalog.elra.info/

the data was treated or processed. It is possible, that most of them anonymized their data, but it is not clearly stated. Other data collected was generally written data such as news (37%), spoken data (11%) and annotations (27%).

In LREC a considerable amount of data from the medical domain, recordings of elderly, pathological voices and data from proficiency observations, such as children or foreign language learner was reported (7%). But in only 10% of the cases anonymization was reported or became obvious through the webpage or published pictures.

5 Suggestions for future direction

From the above presented analysis, we raise several discussion points, which the NLP community should address together. The following is meant as a starting point to flesh out a *code of conduct* and potential future activities to improve the situation.

Data Collection and Usage This addresses issues such as how to collect data, how to pre-/post-process data (i.e. anonymization) and recommendations for available tools supporting these. Additionally, guidelines on how to present data in publications and presentations should enforce anonymization. This could be supported by allowing one additional page for submitted papers, where details on collections, procedures and treatement are given. A checklist both for **authors** and **reviewers** should contain at least:

- Has data been collected?
- How was this data collected and processed?
- Was previously available data used/extended – which one?
- Is a link or a contact given?
- Where does it point (private page, research institute, official repository)?

For journals the availability and usability of data (and potentially code) should be mandatory, similar to Nature and PLOS (see Section 2).

Data Distribution This addresses issues on how data should be distributed to the community, respecting data privacy issues as well. We should define standards for publications that are not tied to a specific lab or even the personal website of a researcher, similar to recommended repositories for Nature or PLOS (see Section 2), but rather provide means and guidelines to gather, work with and publish data. On publication, a defined set of meta data should be provided. These should also include information on methods and tools, which have been used to process the data. This simplifies the reproduction of experiments and results.[19] All of this could be collected in a repository, where code and data is stored. Various efforts in this direction already exist, such as LRE Map[20] or the ACL Data and Code Repository[21]. The ACL Repository currently lists only 9 resources from 2008 to 2011. The LRE Map contains over 2,000 corpora, but the newest dates from LREC 2014. So the data that was analyzed here, has not been provided there.

Adding a reproducibility section to conferences and journals in the NLP domain would support the validation of previously presented results. Studies verified by independent researchers could be raised in the awareness and given appropriate credit to both original researchers and the verification. This could be tied together with extending, encouraging, enforcing the usage of data repositories such as the ACL Repository or the LRE Map and find common interfaces between the various efforts. On the long term, virtual research environments would allow for working with sensitive data without distributing it, which would foster the collaboration across research labs.

6 Future Work

Future work includes extending this preliminary study in two directions: earlier publications and how usable are published data sets. Are various high-profile studies actually replicable and what can we learn from the results?

Additionally, the suggestions sketched in the previous section have to be fleshed out and put to action in a continious revision process.

Acknowledgments

This work was partially supported by the DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1). We would like to thank the reviewers for their valuable comments that helped to considerably improve the paper.

[19] Ideally, a labbook or experiment protocol containing all the necessary information about the experiments should be published as well.

[20] http://www.resourcebook.eu/

[21] https://www.aclweb.org/aclwiki/index.php?title=ACL_Data_and_Code_Repository

References

Julian Bleicken, Thomas Hanke, Uta Salden, and Sven Wagner. 2016. Using a Language Technology Infrastructure for German in order to Anonymize German Sign Language Corpus Data. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

António Branco, Nicoletta Calzolari, and Khalid Choukri, editors. 2016. Portorož, Slovenia. An LREC 2016 Workshop.

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hlne Mazo, Asuncin Moreno, Jan Odijk, and Stelios Piperidis., editors. 2016. *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association, Portorož, Slovenia, May 23–28, 2016. published online at: http://www.lrec-conf.org/proceedings/lrec2016/index.html.

Kevin Cohen, Jingbo Xia, Christophe Roeder, and Lawrence Hunter. 2016. Reproducibility in Natural Language Processing: A Case Study of two R Libraries for Mining PubMed/MEDLINE. In *4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 6–12, Portorož, Slovenia, May. An LREC 2016 Workshop.

Christian Collberg, Todd Proebsting, and Alex M. Warren. 2015. Repeatability and Benefaction in Computer Systems Research – A Study and a Modest Proposal. Technical Report TR 14-04, University of Arizona.

Jana Diesner and Chieh-Li Chin. 2016. Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data. In *ETHI-CA2 2016: ETHics In Corpus Collection, Annotation & Application*, Portorož, Slovenia, May. An LREC 2016 Workshop.

Katrin Erk and Noah A. Smith, editors. 2016. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, August.

Antske Fokkens, Marieke van 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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria, August. Association for Computational Linguistics.

Karën Fort and Alain Couillault. 2016. Yes, We Care! Results of the Ethics and Natural Language Processing Surveys. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Gil Francopoulo, Joseph Mariani, and Patrick Paroubek. 2016. Linking Language Resources and NLP Papers. In *4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 24–32, Portorož, Slovenia, May. An LREC 2016 Workshop.

Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Gabriella Pardelli, Irene Russo, Roberto Bartolini, Fahad Khan, Claudia Soria, and Nicoletta Calzolari. 2016. LREC as a Graph: People and Resources in a Network. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August. Association for Computational Linguistics.

Elizabeth Iorns. 2012. Is medical science built on shaky foundations? `https://www.newscientist.com/article/mg21528826.000-is-medical-science-built-on-shaky-foundations/`, September.

Val Jones. 2009. Science-Based Medicine 101: Reproducibility. `https://sciencebasedmedicine.org/science-based-medicine-101-reproducibility/`, August.

Kevin Knight, Ani Nenkova, and Owen Rambow, editors. 2016. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, June.

Yuji Matsumoto and Rashmi Prasad, editors. 2016. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, December.

Jozef Milšutka, Ondřej Košarko, and Amir Kamran. 2016. SHORTREF.ORG – Making URLs Easy-to-Cite. In *4REAL Workshop: Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language*, pages 19–23, Portorož, Slovenia, May. An LREC 2016 Workshop.

Limor Peer, Ann Green, and Elizabeth Stephenson. 2014. Committing to Data Quality Review. *International Journal of Digital Curation*, 9(1):263–291.

Limor Peer. 2014. Mind the gap in data reuse: Sharing data is necessary but not sufficient for future reuse. `http://blogs.lse.ac.uk/impactofsocialsciences/2014/03/28/mind-the-gap-in-data-reuse/`, March.

Peter Schaar. 2007. Opinion 4/2007 on the concept of personal data. `http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf`.

Information Services and Technology. 2009. Sensitive Data: Your Money AND Your Life. `http://web.mit.edu/infoprotect/docs/protectingdata.pdf`, January.

Jian Su, Kevin Duh, and Xavier Carreras, editors. 2016. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, November.

Stephen Wu, Tamara Timmons, Amy Yates, Meikun Wang, Steven Bedrick, William Hersh, and Hongfang Liu. 2016. On Developing Resources for Patient-level Information Retrieval. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).