A Systematic Review of Reproducibility Research in Natural Language Processing

Anya Belz
University of Brighton, UK
a.s.belz@brighton.ac.uk

Shubham Agarwal
Heriot Watt University, UK
sa201@hw.ac.uk

Anastasia Shimorina
Université de Lorraine / LORIA, France
anastasia.shimorina@loria.fr

Ehud Reiter
University of Aberdeen, UK
e.reiter@abdn.ac.uk

Abstract

Against the background of what has been termed a reproducibility crisis in science, the NLP field is becoming increasingly interested in, and conscientious about, the reproducibility of its results. The past few years have seen an impressive range of new initiatives, events and active research in the area. However, the field is far from reaching a consensus about how reproducibility should be defined, measured and addressed, with diversity of views currently increasing rather than converging. With this focused contribution, we aim to provide a wide-angle, and as near as possible complete, snapshot of current work on reproducibility in NLP, delineating differences and similarities, and providing pointers to common denominators.

1 Introduction

Reproducibility is one of the cornerstones of scientific research: inability to reproduce results is, with few exceptions, seen as casting doubt on their validity. Yet it is surprisingly hard to achieve, 70% of scientists reporting failure to reproduce someone else's results, and more than half reporting failure to reproduce their own, a state of affairs that has been termed the 'reproducibility crisis' in science (Baker, 2016). Following a history of troubling evidence regarding difficulties in reproducing results (Pedersen, 2008; Mieskes et al., 2019), where 24.9% of attempts to reproduce own results, and 56.7% of attempts to reproduce another team's results, are reported to fail to reach the same conclusions (Mieskes et al., 2019), the machine learning (ML) and natural language processing (NLP) fields have recently made great strides towards recognising the importance of, and addressing the challenges posed by, reproducibility: there have been several workshops on reproducibility in ML/NLP including the Reproducibility in ML Workshop at ICML'17, ICML'18 and ICLR'19; the Reproducibility Challenge at ICLR'18, ICLR'19, NeurIPS'19, and NeurIPS'20; LREC'20 had a reproducibility track and shared task (Branco et al., 2020); and NeurIPS'19 had a reproducibility programme comprising a code submission policy, a reproducibility challenge for ML results, and the ML Reproducibility checklist (Whitaker, 2017), later also adopted by EMNLP'20 and AAAI'21. Other conferences have foregrounded reproducibility via calls, chairs' blogs,[1] special themes and social media posts. Sharing code, data and supplementary material providing details about data, systems and training regimes[2] is firmly established in the ML/NLP community, virtually all main events now encouraging and making space for it. Reproducibility even seems set to become a standard part of reviewing processes via checklists. Far from beginning to converge in terms of standards, terminology and underlying definitions, however, this flurry of work is currently characterised by growing diversity in all these respects. We start below by surveying concepts and definitions in reproducibility research, areas of particular disagreement, and identify categories of work in current NLP reproducibility research. We then use the latter to structure the remainder of the paper.

Selection of Papers: We conducted a structured review employing a stated systematic process for identifying all papers in the field that met specific criteria. Structured reviews are a type of meta-review more common in fields like medicine but beginning to be used more in NLP (Reiter, 2018; Howcroft et al., 2020).

Specifically, we selected papers as follows. We

[1] https://2020.emnlp.org/blog/2020-05-20-reproducibility

[2] There are some situations where it is difficult to share data, e.g. because the data is commercially confidential or because it contains sensitive personal information. But the increasing expectation in NLP is that authors should share as much as possible, and justify cases where it is not possible.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 381–393
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

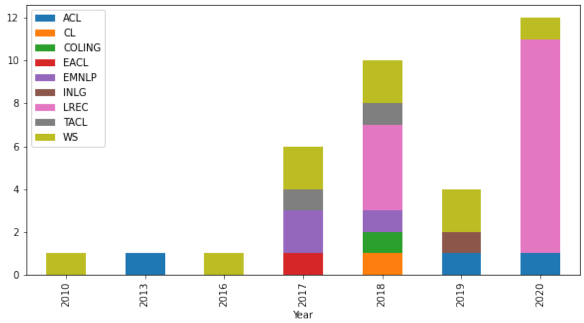

Figure 1: 35 papers from ACL Anthology search, by year and venue.[5]

searched the ACL Anthology for titles containing either of the character strings *reproduc* or *replica*, either capitalised or not.[3] This yielded 47 papers; following inspection we excluded 12 of the papers as not being about reproducibility in the present sense.[4] We found 25 additional papers in non-ACL NLP/ML sources, and a further 7 in other fields.

Figure 1 shows[5] how the 35 papers from the ACL Anthology search are distributed over years: one paper a year at most until 2017/18 when interest seems to have increased spontaneously, before dropping off again. The renewed high numbers for 2020 are almost entirely due to the LREC REPROLANG shared task (see Section 5 below).

2 Terminology and Frameworks

Reproducibility research in NLP and beyond uses a bewildering range of closely related terms, often with conflicting meaning, including reproducibility, repeatability, replicability, recreation, re-run, robustness, repetition and generalisability. The fact that no formal definition of any of these terms singly, let alone in relation to each other, is generally accepted as standard, or even predominant, in NLP at present, is clearly a problem for a survey paper. In this section, we review usage before identifying common-ground terminology that will enable us to talk about the research we survey.

The two most frequently used 'R-terms', *reproducibility* and *replicability*, are also the most problematic. For the ACM (Association for Computing Machinery, 2020), results have been *reproduced* if "obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author," and *replicated* if "obtained in a subsequent study by a person or team other than the authors, without the use of author-supplied artifacts" (although initially the terms were defined the other way around[6]). The definitions are tied to team and software (artifacts), but it is unclear how much of the latter have to be the same for reproducibility, and how different the team needs to be for either concept.

Rougier et al. (2017) tie definitions (just) to new vs. original software: "*Reproducing* the result of a computation means running the same software on the same input data and obtaining the same results. [...] *Replicating* a published result means writing and then running new software based on the description of a computational model or method provided in the original publication, and obtaining results that are similar enough to be considered equivalent." It is clear from the many reports of failures to obtain 'same results' with 'same software and data' in recent years that the above definitions raise practical questions such as how to tell 'same software' from 'new software,' and how to determine equivalence of results.

Wieling et al. (2018) define *reproducibility* as "the exact re-creation of the results reported in a publication using the same data and methods," but then discuss (failing to) *replicate* results without defining that term, while also referring to the "unfortunate swap" of the definitions of the two terms put forward by Drummond (2009).

Whitaker (2017), followed by Schloss (2018), tie definitions to data as well as code:

		Data	
		Same	*Different*
Code	*Same*	Reproducible	Replicable
	Different	Robust	Generalisable

The different definitions of *reproducibility* and *replicability* above, put forward in six different contexts, are not compatible with each other. Grappling with degrees of similarity between properties of experiments such as the team, data and software involved, and between results obtained, each draws the lines between terms differently, and moreover variously treats reproducibility and replicability as properties of either systems or results. All are patchy, not accounting for some circumstances, e.g. a team reproducing its own results, not defining

[3] `grep -E ‘title *=.*(r|R)\}*(eproduc|eplica)’ anthology.bib`

[4] Most were about annotation and data replication.

[5] Data and code: https://github.com/shubhamagarwal92/eacl-reproducibility

[6] ACM swapped definitions of the two terms when prompted by NISO to "harmonize its terminology and definitions with those used in the broader scientific research community." (Association for Computing Machinery, 2020).

some concepts, e.g. sameness, or not specifying what can serve as a 'result,' e.g. leaving the status of human evaluations and dataset recreations unclear.

The extreme precision of the definitions of the International Vocabulary of Metrology (VIM) (JCGM, 2012) (which the ACM definitions are supposed to be based on but aren't quite) offers a common terminological denominator. The VIM definitions of *reproducibility* and *repeatability* (no other R-terms are defined) are entirely general, made possible by two key differences compared to the NLP/ML definitions above. Firstly, in a key conceptual shift, reproducibility and repeatability are properties of **measurements** (not of systems or abstract findings). The important difference is that the concept of reproducibility now references a specified way of obtaining a **measured quantity value** (which can be an evaluation metric, statistical measure, human evaluation method, etc. in NLP). Secondly, reproducibility and repeatability are defined as the *precision* of a measurement under specified conditions, i.e. the distribution of the quantity values obtained in *repeat* (or *replicate*) measurements.

In VIM, **repeatability** is the precision of measurements of the same or similar object obtained under the same conditions, as captured by a specified set of **repeatability conditions**, whereas **reproducibility** is the precision of measurements of the same or similar object obtained under different conditions, as captured by a specified set of **reproducibility conditions**. See Appendix C for a full set of VIM definitions of the bold terms above.

To make the VIM terms more recognisable in an NLP context, we also call repeatability **reproducibility under same conditions**, and (VIM) reproducibility **reproducibilty under varied conditions**. Finally, we refer to experiments carrying out repeat measurements regardless of same/varied conditions as 'reproduction studies.'

Categories of Reproducibility Research: Using the definitions above, the work we review in the remainder of the paper falls into three categories (corresponding to Sections 3–5):

Reproduction under same conditions: As near as possible exact recreation or reuse of an existing system and evaluation set-up, and comparison of results.[7]

Reproduction under varied conditions: Reproduction studies with deliberate variation of one or more aspects of system and/or measurement, and comparison of results.

Multi-test studies: Multiple reproduction studies connected e.g. because of an overall multi-test design, and/or use of same methodology.

3 Reproduction Under Same Conditions

Papers reporting reproductions under same conditions account for the bulk of NLP reproducibility research to date. The difficulty of achieving 'sameness of system' has taken up a lot of the discussion space. As stressed by many papers (Crane, 2018; Millour et al., 2020), recreation attempts have to have access to code, data, full details/assumptions of algorithms, parameter settings, software and dataset versions, initialisation details, random seeds, run-time environment, hardware specifications, etc.

A related and striking finding, confirmed by multiple repeatability studies, is that results often depend in surprising ways and to surprising degrees on seemingly small differences in model parameters and settings, such as rare-word thresholds, treatment of ties, or case normalisation (Fokkens et al., 2013; Van Erp and Van der Meij, 2013; Dakota and Kübler, 2017). Effects are often discovered during system recreation from incomplete information, when a range of values is tested for missing details. The concern is that the ease with which such NLP results are perturbed casts doubt on their generalisability and robustness.

The difficulties in recreating, or even just rerunning, systems with same results have led to growing numbers of reproducibility checklists (Olorisade et al., 2017; Pineau, 2020), and tips for making system recreation easier, e.g. the PyTorch (Paszke et al., 2017) recommended settings.[8]

We analysed reproduction studies under same conditions from 34 pairs of papers, and identified 549 individual score pairs where reproduction object, method and outcome were clear enough to include in comparisons (for a small number of papers this meant excluding some scores). Table 1 in Appendix A provides a summary of the results. In 36 cases, the reproduction study did not produce scores, e.g. because resource limits were reached,

[7] This excludes the countless cases where results for a previous method are used as a baseline or other comparitor, but experiments are not run again.

[8] https://pytorch.org/docs/stable/notes/randomness.html

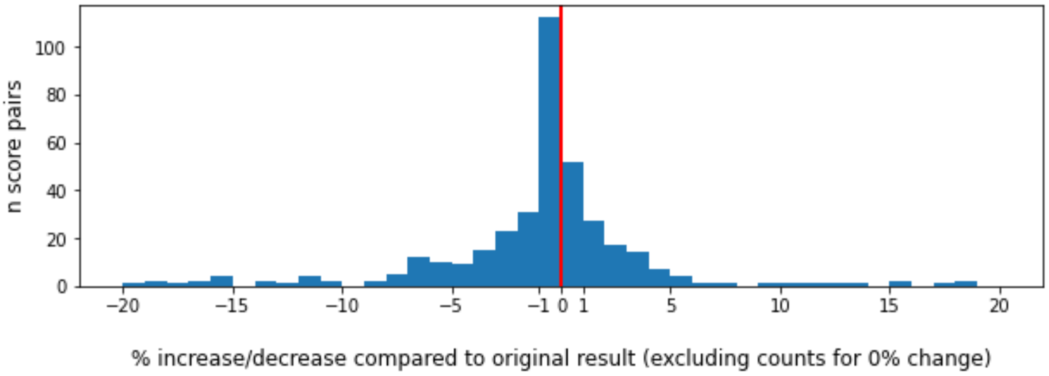

Figure 2: Histogram of percentage differences between original and reproduction scores (bin width = 1; clipped to range -20..20).

or code didn't work. This left 513 cases where the reproduction study produced a value that could be compared to the original score. Out of these, just 77 (14.03%) were exactly the same. Out of the remaining 436 score pairs, in 178 cases (40.8%), the reproduction score was better than the original, and in 258 cases (59.2%) it was worse.

We also examined the size of the differences between original and reproduction scores. For this purpose we computed percentage change (increase or decrease) for each score pair, and looked at the distribution of size and direction of change. For this analysis, we excluded score pairs where one or both scores were 0, as well as 4 very large outliers (all increases). Results are shown in the form of a histogram with bin width 1 (and clipped to range -20..20) in Figure 2. The plot clearly shows the imbalance between worse (60% of non-same scores) and better (40%) reproduction scores. The figure also shows that a large number of differences fall in the -1..1 range. However, the majority of differences, or 3/5, are greater than +/-1%, and about 1/4 are greater than +/-5%.

4 Reproduction Under Varied Conditions

Reproduction studies under varied conditions deliberately vary one or more aspects of system, data or evaluation in order to explore if similar results can be obtained. There are far fewer papers of this type (see Table 2 in Appendix B for an overview) than papers reporting reproduction studies under same conditions; however, note that we are not including papers here that use an existing method for a new language, dataset or domain, without controlling for other conditions being the same in experiments.

Horsmann and Zesch (2017) pick up strong results by Plank et al. (2016) showing LSTM tagging to outperform CRF and HMM taggers, and test whether they can be confirmed for datasets with finer-grained tag sets. Using 27 corpora (21 languages) with finer-grained tag sets, they systematically compare results for the 3 models, and show that LSTMs do perform better, and that their superiority grows in proportion to tag set size.

Htut et al. (2018a) recreate the grammar induction model PRPN (Shen et al., 2018), testing different versions with different data. PRPN is confirmed to be “strikingly effective” at latent tree learning. In a subsequent repeat study under same conditions, Htut et al. (2018b) test PRPN using the authors' own code, obtaining the same headline result.

Millour et al. (2020) attempt to get the POS tagger for Alsatian from Magistry et al. (2018) to work with the same accuracy for a different dataset. Collaborating with, and using resources provided by, the original authors and recreating some where necessary, the best result obtained was 0.87 compared to the original 0.91.

Abdellatif and Elgammal (2020) varied conditions of reproduction for classification results by Howard and Ruder (2018), and were able to confirm outcomes for three new non-English datasets, in all three respects (value, finding, conclusion) identified by Cohen et al. (2018).

Pluciński et al. (2020) and Garneau et al. (2020) both find that the cross-lingual word embedding mappings proposed by Artetxe et al. (2018) yield worse results on more distant language pairs.

Vajjala and Rama (2018)'s automatic essay scoring classification system was tested on different datasets and/or languages in three studies (Arhiliuc et al., 2020; Caines and Buttery, 2020; Huber and Çöltekin, 2020) all of which found performance to drop on the new data.

5 Multi-test and Multi-lab Studies

Work in this category is *multi-test*, in the sense of involving multiple reproduction studies, in a uniform framework using uniform methods. Some of it is also *multi-lab* in that reproduction studies are carried out by more than one research team. For example, in one multi-test repeatability study, Wieling et al. (2018) randomly select five papers each from ACL'11 and ACL'16 for which code/data was available. In a uniform design, original authors were contacted for help if needed, a maximum time limit of 8h was imposed, and all work was done by the same Masters student. It's not clear how scores were selected (not all are attempted), and reasons for failure are not always clear even from linked material. Of the 120 score pairs obtained, 60 were the same, 12 reproduction scores were better, 22 were worse, and 26 runs failed (including exceeding the time limit). See Table 1 for summary.

Rodrigues et al. (2020) recreated six SemEval'18 systems from the Argument Reasoning Comprehension Task, following system descriptions and/or reusing code, with no time limit. Scores were the same for one system, and within +/- 0.036 points for the other five; the SemEval ranking was exactly the same. Systems were also run on a corrected version of the shared-task data (which contained unwitting clues). This resulted in much lower scores casting doubt on the validity of the original shared task results.

REPROLANG (Branco et al., 2020) is so far the only multi-lab (as well as multi-test) study of reproducibility in NLP. It was run as a selective shared task, and required participants to conform to uniform rules. 11 papers were selected for reproduction via an open call and direct selection. Participants had to 'reproduce the paper,' using information contained/linked in it. Participants submitted (a) a report on the reproduction, and (b) the software used to obtain the results as a Docker container (controlling variation from dependencies and run-time environments) on GitLab. Submissions were reviewed in great detail, submitted code was test-run and checked for hard-coding of results. 11 out of 18 submissions were judged to conform with requirements. One original paper (Vajjala and Rama, 2018) attracted four reproductions (Bestgen, 2020; Huber and Çöltekin, 2020; Caines and Buttery, 2020; Arhiliuc et al., 2020) in what must be a groundbreaking first in NLP. See Table 1 for summaries of all 11 reproductions. An aspect the organisers did not control was how to draw conclusions about reproducibility; most contributions provide binary conclusions but vary in how similar they require results to be for success. E.g. the four papers reproducing Vajjala and Rama (2018) all report similarly large deviations, but only one (Arhiliuc et al., 2020) concludes that the reproduction was not a success.

6 Conclusions

It seemed so simple: share all data, code and parameter settings, and other researchers will be able to obtain the same results. Yet the systems we create remain stubbornly resistant to this goal: a tiny 14.03% of the 513 original/reproduction score pairs we looked at were the same. At the same time, worryingly small differences in code have been found to result in big differences in performance.

Another striking finding is that reproduction under same conditions far more frequently yields results that are worse than results that are better: we found 258 out of 436 non-same reproduction results (59.2%) to be worse, echoing findings from psychology (Open Science Collaboration, 2015). Why this should be the case for reproduction under *same* conditions is unclear. It is probably to be expected for reproduction under *different* conditions, as larger parameter spaces, more datasets and languages etc., are tested subsequently, and the original work may have selected better results.

There is a lot of work going on in NLP on reproducibility right now; it is to be hoped that we can solve the vexing and scientifically uninteresting problem of how to rerun code and get the same results soon, and move on to addressing far more interesting questions of how reliable, stable and generalisable promising NLP results really are.

Acknowledgments

We thank our reviewers for their valuable feedback. Shubham Agarwal's PhD fees are supported by Adeptmind Inc., Toronto, Canada.

References

Mohamed Abdellatif and Ahmed Elgammal. 2020. ULMFiT replication. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5579–5587, Marseille, France. European Language Resources Association.

Cristina Arhiliuc, Jelena Mitrović, and Michael Granitzer. 2020. Language proficiency scoring. In *Pro-*

ceedings of The 12th Language Resources and Evaluation Conference, pages 5624–5630, Marseille, France. European Language Resources Association.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Association for Computing Machinery. 2020. Artifact review and badging Version 1.1, August 24, 2020. `https://www.acm.org/publications/policies/artifact-review-and-badging-current`.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760.

Monya Baker. 2016. Reproducibility crisis. *Nature*, 533(26):353–66.

Yves Bestgen. 2020. Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France. European Language Resources Association.

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

Bernd Bohnet, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez. 2018. Morphosyntactic tagging with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2642–2652, Melbourne, Australia. Association for Computational Linguistics.

S.R.K. Branavan, 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 Association for Computational Linguistics: Human Language Technologies*, pages 268–277, Portland, Oregon, USA. Association for Computational Linguistics.

António Branco, Nicoletta Calzolari, Piek Vossen, Gertjan Van Noord, Dieter van Uytvanck, João Silva, Luís Gomes, André Moreira, and Willem Elbers. 2020. A shared task of a new, collaborative type to foster reproducibility: A first exercise in the area of language science and technology with REPROLANG2020. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5539–5545, Marseille, France. European Language Resources Association.

Ana Brassard, Tin Kuculo, Filip Boltužić, and Jan Šnajder. 2018. TakeLab at SemEval-2018 task12: Argument reasoning comprehension with skip-thought vectors. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1133–1136, New Orleans, Louisiana. Association for Computational Linguistics.

Andrew Caines and Paula Buttery. 2020. REPROLANG 2020: Automatic proficiency scoring of Czech, English, German, Italian, and Spanish learner essays. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5614–5623, Marseille, France. European Language Resources Association.

HongSeok Choi and Hyunju Lee. 2018. GIST at SemEval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777, New Orleans, Louisiana. Association for Computational Linguistics.

Maximin Coavoux and Benoît Crabbé. 2016. Neural greedy constituent parsing with dynamic oracles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 172–182, Berlin, Germany. Association for Computational Linguistics.

K Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany J Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin, and Lawrence E Hunter. 2018. Three dimensions of reproducibility in natural language processing. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2018, page 156. NIH Public Access.

Michael Collins. 1999. Head-driven statistical models for natural language parsing. Ph.D. thesis.

Michael Cooper and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5588–5594, Marseille, France. European Language Resources Association.

Matt Crane. 2018. Questionable answers in question answering research: Reproducibility and variability of published results. *Transactions of the Association for Computational Linguistics*, 6:241–252.

Daniel Dakota and Sandra Kübler. 2017. Towards replicability in parsing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 185–194, Varna, Bulgaria. INCOMA Ltd.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. Presented at 4th Workshop on Evaluation Methods for Machine Learning held at ICML'09.

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. Association for Computational Linguistics.

Paula Fortuna, Juan Soler-Company, and Sérgio Nunes. 2019. Stop PropagHate at SemEval-2019 tasks 5 and 6: Are abusive language classification results reproducible? In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 745–752, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Nuno Freire, José Borbinha, and Pável Calado. 2012. An approach for named entity recognition in poorly structured data. In *Extended Semantic Web Conference*, pages 718–732. Springer.

Qiaozi Gao, 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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1814–1824, Berlin, Germany. Association for Computational Linguistics.

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand, and Luc Lamontagne. 2020. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings: Making the method robustly reproducible as well. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5546–5554, Marseille, France. European Language Resources Association.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Yulan He, 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 Association for Computational Linguistics: Human Language Technologies*, pages 123–131, Portland, Oregon, USA. Association for Computational Linguistics.

Tobias Horsmann and Torsten Zesch. 2017. Do LSTMs really work so well for PoS tagging? – a replication study. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 727–736, Copenhagen, Denmark. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Phu Mon Htut, Kyunghyun Cho, and Samuel Bowman. 2018a. Grammar induction with neural language models: An unusual replication. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4998–5003, Brussels, Belgium. Association for Computational Linguistics.

Phu Mon Htut, Kyunghyun Cho, and Samuel Bowman. 2018b. Grammar induction with neural language models: An unusual replication. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 371–373, Brussels, Belgium. Association for Computational Linguistics.

Zhiting Hu, 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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2410–2420, Berlin, Germany. Association for Computational Linguistics.

Eva Huber and Çağrı Çöltekin. 2020. Reproduction and replication: A case study with automatic essay scoring. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5603–5613, Marseille, France. European Language Resources Association.

JCGM. 2012. International vocabulary of metrology-basic and general concepts and associated terms (vim).

Yung Han Khoe. 2020. Reproducing a morphosyntactic tagger with a meta-BiLSTM model over context sensitive token encodings. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5563–5568, Marseille, France. European Language Resources Association.

Taeuk Kim, Jihun Choi, and Sang-goo Lee. 2018. SNU_IDS at SemEval-2018 task 12: Sentence encoder with contextualized vectors for argument reasoning comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1083–1088, New Orleans, Louisiana. Association for Computational Linguistics.

Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 590–599, Portland, Oregon, USA. Association for Computational Linguistics.

Pierre Magistry, Anne-Laure Ligozat, and Sophie Rosset. 2018. Étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux. In *Conférence sur le Traitement Automatique des Langues Naturelles*, Rennes, France.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 768–775, Varna, Bulgaria. INCOMA Ltd.

Alice Millour, Karën Fort, and Pierre Magistry. 2020. Répliquer et étendre pour l'alsacien "étiquetage en parties du discours de langues peu dotées par spécialisation des plongements lexicaux" (replicating and extending for Alsatian : "POS tagging for low-resource languages by adapting word embeddings"). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 29–37, Nancy, France. ATALA et AFCP.

Andrew Moore and Paul Rayson. 2018. Bringing replication and reproduction together with generalisability in NLP: Three reproduction studies for target dependent sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1132–1144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Preslav Nakov and Hwee Tou Ng. 2011. Translating from morphologically complex languages: A paraphrase-based approach. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1298–1307, Portland, Oregon, USA. Association for Computational Linguistics.

Garrett Nicolai and Grzegorz Kondrak. 2016. Leveraging inflection tables for stemming and lemmatization. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1138–1147, Berlin, Germany. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2018. NLITrans at SemEval-2018 task 12: Transfer of semantic knowledge for argument comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1099–1103, New Orleans, Louisiana. Association for Computational Linguistics.

Babatunde Kazeem Olorisade, Pearl Brereton, and Peter Andras. 2017. Reproducibility of studies on text mining for citation screening in systematic reviews: evaluation and checklist. *Journal of biomedical informatics*, 73:1–13.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS-W*.

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

Joelle Pineau. 2020. The machine learning reproducibility checklist v2.0.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418, Berlin, Germany. Association for Computational Linguistics.

Kamil Pluciński, Mateusz Lango, and Michał Zimniewicz. 2020. A closer look on unsupervised cross-lingual word embeddings mapping. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5555–5562, Marseille, France. European Language Resources Association.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Kyeongmin Rim, Jingxuan Tu, Kelley Lynch, and James Pustejovsky. 2020. Reproducing neural ensemble classifier for semantic relation extraction in-Scientific papers. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5569–5578, Marseille, France. European Language Resources Association.

João Rodrigues, Ruben Branco, João Silva, and António Branco. 2020. Reproduction and revival of the argument reasoning comprehension task. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5055–5064, Marseille, France. European Language Resources Association.

Jonathan Rotsztejn, Nora Hollenstein, and Ce Zhang. 2018. ETH-DS3Lab at SemEval-2018 task 7: Effectively combining recurrent and convolutional neural networks for relation classification and extraction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 689–696, New Orleans, Louisiana. Association for Computational Linguistics.

Nicolas P Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A Barba, Fabien CY Benureau, C Titus Brown, Pierre De Buyl, Ozan Caglayan, Andrew P Davison, et al. 2017. Sustainable computational science: the rescience initiative. *PeerJ Computer Science*, 3:e142.

Christina Sauper, Aria Haghighi, and Regina Barzilay. 2011. Content models with attitude. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 350–358, Portland, Oregon, USA. Association for Computational Linguistics.

Patrick D Schloss. 2018. Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio*, 9(3).

Yikang Shen, Zhouhan Lin, Chin wei Huang, and Aaron Courville. 2018. Neural language modeling by jointly learning syntax and lexicon. In *International Conference on Learning Representations*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307.

Junfeng Tian, Man Lan, and Yuanbin Wu. 2018. ECNU at SemEval-2018 task 12: An end-to-end attention-based neural network for the argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1094–1098, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Tian, Naoaki Okazaki, and Kentaro Inui. 2016. Learning semantically and additively compositional distributional representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1277–1287, Berlin, Germany. Association for Computational Linguistics.

Sowmya Vajjala and Taraka Rama. 2018. Experiments with universal CEFR classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 147–153, New Orleans, Louisiana. Association for Computational Linguistics.

Marieke Van Erp and Lourens Van der Meij. 2013. Reusable research? a case study in named entity recognition. Technical report.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Bo Wang, Maria Liakata, Arkaitz Zubiaga, and Rob Procter. 2017. Tdparse: Multi-target-specific sentiment recognition on twitter. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 483–493.

Kirstie Whitaker. 2017. The MT Reproducibility Checklist. `https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf`.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Meiqian Zhao, Chunhua Liu, Lu Liu, Yan Zhao, and Dong Yu. 2018. BLCU_NLP at SemEval-2018 task 12: An ensemble model for argument reasoning based on hierarchical attention. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1104–1108, New Orleans, Louisiana. Association for Computational Linguistics.

Appendices

A Table of Reproductions Under Same Conditions

Original paper	Reproduction study (same conditions)	NLP task	Summary of score differences
Collins (1999)	Gildea (2001)	Phrase-structure parsing	+16.7% error on Model 1 results
Collins (1999)	Bikel (2004)	Phrase-structure parsing	+11% error on Model 2 results on WSJ00; later same results with help from Collins
Freire et al. (2012)	Van Erp and Van der Meij (2013)	NER	"Despite feedback from Freire [...], results remained 20 points below those reported in Freire et al. (2012) in overall F-score" (Fokkens et al., 2013)
Nakov and Ng (2011)	Wieling et al. (2018)	MT	*Unsuccessful (scripts did not work)
He et al. (2011)	Wieling et al. (2018)	Sentiment analysis	*-0.18 points
Sauper et al. (2011)	Wieling et al. (2018)	Topic modelling	*Unsuccessful on 3 scores (8h cut-off reached)
Liang et al. (2011)	Wieling et al. (2018)	Question answering	*Exact reproduction of 2 scores in 4h
Branavan et al. (2011)	Wieling et al. (2018)	Joint learning of game strategy and text selection from game manual	*Unsuccessful on 7 scores (scripts did not generate output)
Coavoux and Crabbé (2016)	Wieling et al. (2018)	Constituent parsing	*9/18 scores same, 9/18 parser did not complete for 4 languages
Gao et al. (2016)	Wieling et al. (2018)	Semantic role grounding	*Exact reproduction of 44/72 scores, 17 worse, 11 better, average -0.62 points
Hu et al. (2016)	Wieling et al. (2018)	Sentiment analysis, NER	*exact reproduction of 1/2 scores, 1 worse -0.2 points
Nicolai and Kondrak (2016)	Wieling et al. (2018)	Stemming, lemmatisation	*2/8 scores -3.4 and -1.55 points, 6/8 scores took longer than 8h cut-off
Tian et al. (2016)	Wieling et al. (2018)	Sentence completion	*4/6 scores reproduced exactly, 2/6 differed -0.1 and +0.01 %-points).
Badjatiya et al. (2017)	Fortuna et al. (2019)	hate speech detection	reproduction under same conditions not possible due to issue with code; recreated/corrected system did well at OffensEval'19 but not at HatEval'19
Choi and Lee (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	1/1 score +0.002 points
Zhao et al. (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	1/1 score +0.038 points
Tian et al. (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	1/1 score -0.021 points
Niven and Kao (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	1/1 score +0.033 points
Kim et al. (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	1/1 score -0.022 points
Brassard et al. (2018)	Rodrigues et al. (2020)	Argument Reasoning Comprehension Task	Exact reproduction of 1/1 score.
Artetxe et al. (2018)	Garneau et al. (2020)	Cross-lingual Mappings of Word Embeddings	Main scores: 2/8 same, 1/8 -0.1, 5/8 +0.1 to +0.3; ablation scores: 4/40 scores same, 19/40 +0.1 to 6.9, 9/40 -0.1 to -0.9, 8/40 took too long
Artetxe et al. (2018)	Pluciński et al. (2020)	Cross-lingual Mappings of Word Embeddings	Main scores: 10/14 better, 4/14 worse; ablation scores: 3/48 scores same, 31/48 better, 14/48 worse
Bohnet et al. (2018)	Khoe (2020)	POS and morphological tagging	POS tagging scores: 35/41 worse, 6/41 better; morph. tagging: 43/46 worse, 3/46 better
Rotsztejn et al. (2018)	Rim et al. (2020)	Relation extraction and classification (SemEval'18 T7)	4 subtasks: 4/4 scores worse, up to 9.04 points; subtask 1.1 by relation: 3/6 worse, 3/6 better
Nisioi et al. (2017)	Cooper and Shardlow (2020)	Simplification	NTS default system: 1/2 automatic scores better, 1/2 automatic scores worse; 2/2 human scores worse

Table continued on next page.

Original paper	Reproduction study (recreation of system)	NLP task	Summary of score differences
Vajjala and Rama (2018)	Bestgen (2020)	Automatic essay scoring (classification)	multilingual: 6/11 better, 5/11 worse; monolingual: 15/27 better, 11/27 worse, 1/27 same; crosslingual: 5/8 better, 1/8 worse, 2/8 same
Vajjala and Rama (2018)	Huber and Çöltekin (2020)	Automatic essay scoring (classification)	multilingual: 3/11 better, 8/11 worse; monolingual: 8/27 better, 19/27 worse; crosslingual: 6/8 better, 2/8 worse
Vajjala and Rama (2018)	Caines and Buttery (2020)	Automatic essay scoring (classification)	multilingual: 9/11 better, 2/11 worse; monolingual: 14/27 better, 11/27 worse, 2/27 same; crosslingual: 1/8 better, 7/8 worse
Vajjala and Rama (2018)	Arhiliuc et al. (2020)	Automatic essay scoring (classification)	multilingual: 11/11 worse; monolingual: 7/27 better, 20/27 worse; crosslingual: 1/8 better, 5/8 worse, 2/8 same
Magistry et al. (2018)	Millour et al. (2020)	POS tagging for Alsatian	baseline: same (0.78 Acc); main: worse (Acc 0.87 vs. 0.91)
Howard and Ruder (2018)	Abdellatif and Elgammal (2020)	Sentiment classification, question classification, topic classification	3/6 better, 3/6 worse
Vo and Zhang (2015)	Moore and Rayson (2018)	Target Dependent Sentiment analysis	6/6 better
Wang et al. (2017)	Moore and Rayson (2018)	Target Dependent Sentiment analysis	2/5 better, 3/5 worse
Tang et al. (2016)	Moore and Rayson (2018)	Target Dependent Sentiment analysis	3/3 worse

Table 1: Tabular overview of individual repeatability tests from 34 paper pairs, and a total of 549 score pairs. * = additional information obtained from hyperlinked material.

Where scores obtained in a repeatability study (reproduction under same conditions) are worse than in the original work, this should *not* be interpreted as casting the original work in a negative light. This is because it is normally not possible to create the exact same conditions in repeatability studies (and lower scores can result from such differences), and because the outcome from multiple repeatability studies may be very different.

For a small number of papers, the score pairs included in this table are a subset of scores reported in the paper. More generally, the summary in the last column should not be interpreted as a summary of the whole paper and its findings.

Our intention here is to summarise differences that have been reported in the literature, rather than draw conclusions about what may have caused the differences.

B Table of Reproductions Under Varied Conditions

Original paper	reproduction study (confirmation of finding)	NLP task	Summary of outcome (as interpreted by authors)
Plank et al. (2016)	Horsmann and Zesch (2017)	POS tagging	Confirmed for finer-grained tagsets
Shen et al. (2018)	Htut et al. (2018a,b)	Grammar induction	Overall finding confirmed (that PRPN is a high performing grammar induction method)
Magistry et al. (2018)	Millour et al. (2020)	POS tagging	Not confirmed, reproduction results worse by $>$ 10 BLEU points
Vajjala and Rama (2018)	Arhiliuc et al. (2020)	Automatic essay scoring (classification)	Lower classification results on a corpus of Asian learners' English.
Vajjala and Rama (2018)	Caines and Buttery (2020)	Automatic essay scoring (classification)	Lower classification results for English and Spanish CEFR datasets, and some adversarial data (e.g., scrambled English texts).
Vajjala and Rama (2018)	Huber and Çöltekin (2020)	Automatic essay scoring (classification)	Lower classification results for English Cambridge Learner Corpus.
Artetxe et al. (2018)	Garneau et al. (2020)	Cross-lingual mappings of word embeddings	For other distant language pairs (from English to Estonian, Latvian, Finnish, Persian) the method did not converge or obtained lower scores.
Artetxe et al. (2018)	Pluciński et al. (2020)	Cross-lingual mappings of word embeddings	For other distant language pairs (from English to Czech, Polish) the method did not converge or obtained lower scores.
Howard and Ruder (2018)	**Abdellatif and Elgammal (2020)	Sentiment classification, question classification, topic classification	Confirmed that transfer learning (pretraining) improves final classification accuracy.

Table 2: Tabular overview of individual studies to confirm a previous research finding. * = additional information obtained from hyperlinked material; ** = reproduction study had minor differences, e.g. hyperparameter tuning was omitted (Abdellatif and Elgammal, 2020).

The comments from the caption for Table 1 also apply here, but note that some differences between original and reproduction study are overt and intentional in the case of the papers in this table, whereas they are not intentional and often inadvertent in the case of the papers in Table 1.

C Verbatim VIM and ACM Definitions

2.1 (2.1) **measurement**	process of experimentally obtaining one or more **quantity values** that can reasonably be attributed to a **quantity**
2.15 measurement precision (precision)	closeness of agreement between **indications** or **measured quantity values** obtained by replicate **measurements** on the same or similar objects under specified conditions
2.20 (3.6, Notes 1 and 2) **repeatability condition of measurement** (repeatability condition)	condition of **measurement**, out of a set of conditions that includes the same **measurement procedure**, same operators, same **measuring system**, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time
2.21 (3.6) **measurement repeatability** (repeatability)	**measurement precision** under a set of **repeatability conditions of measurement**
2.24 (3.7, Note 2) **reproducibility condition of measurement** (reproducibility condition)	condition of **measurement**, out of a set of conditions that includes different locations, operators, **measuring systems**, and replicate measurements on the same or similar objects
2.25 (3.7) **measurement reproducibility** (reproducibility)	**measurement precision** under **reproducibility conditions of measurement**
2.3 (2.6) **measurand**	**quantity** intended to be measured

Table 3: VIM definitions of repeatability and reproducibility (JCGM, 2012).

Repeatability (Same team, same experimental setup)	The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
Reproducibility (Different team, same experimental setup)*	The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
Replicability (Different team, different experimental setup)*	The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Results Validated: This badge is applied to papers in which the main results of the paper have been successfully obtained by a person or team other than the author. Two levels are distinguished:

Results Reproduced v1.1 The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the author.

Results Replicated v1.1 The main results of the paper have been independently obtained in a subsequent study by a person or team other than the authors, without the use of author-supplied artifacts.

In each case, exact replication or reproduction of results is not required, or even expected. Instead, the results must be in agreement to within a tolerance deemed acceptable for experiments of the given type. In particular, differences in the results should not change the main claims made in the paper.

Table 4: ACM definitions (bold) and badges (underlined) (Association for Computing Machinery, 2020).