Last Words

Sharing Is Caring: The Future of Shared Tasks

Malvina Nissim[*]
University of Groningen

Lasha Abzianidze[*]
University of Groningen

Kilian Evang[*]
University of Groningen

Rob van der Goot[*]
University of Groningen

Hessel Haagsma[*]
University of Groningen

Barbara Plank[*]
University of Groningen

Martijn Wieling[*]
University of Groningen

1. The Role of Shared Tasks in NLP

Shared tasks are indisputably drivers of progress and interest for problems in NLP. This is reflected by their increasing popularity, as well as by the fact that new shared tasks regularly emerge for under-researched and under-resourced topics, especially at workshops and smaller conferences.

The general procedures and conventions for organizing a shared task have arisen organically over time (Paroubek, Chaudiron, and Hirschman 2007, Section 7). There is no consistent framework that describes how shared tasks should be organized. This is not a harmful thing per se, but we believe that shared tasks, and by extension the field in general, would benefit from some reflection on the existing conventions. This, in turn, could lead to the future harmonization of shared task procedures.

Shared tasks revolve around two aspects: research advancement and competition. We see research advancement as the driving force and main goal behind organizing them. Competition is an instrument to encourage and promote participation. However,

[*] CLCG, Faculty of Arts, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen, The Netherlands.
E-mail: m.nissim@rug.nl.

doi:10.1162/COLI_a_00304

just because these two forces are intrinsic to shared tasks does not mean that they always act in the same direction: Ensuring that the competition is fair is not a necessary requirement for advancing the field, and might even slow down progress.

Our position in this respect is clear: We do believe that (i) advancing the field should be given priority over ensuring fair competition, also because (ii) inequality is partly unsolvable and intrinsic to life. In other words: Equality between competitors is desirable if it does not hinder research advancement.

In the recently established workshop on ethics in NLP,[1] Parra Escartín et al. (2017) raise a set of considerations involving shared tasks, mainly focusing on areas where general ethical concerns regarding good scientific practice intersect with certain aspects of shared tasks. We find that they raise valid concerns, and in this contribution, we address some of them. However, we take a different perspective. Instead of focusing on ethical issues and potential negative effects of the competition aspect, we rather concentrate on how to bolster scientific progress.

We make a simple proposal for the improvement of shared tasks, and discuss how it can help to mitigate the problems raised by Parra Escartín et al. (2017), while not necessarily tackling them directly. We start with assessing the concrete impact and significance of such concerns first.

2. Let's Talk About Problems

Recently, Parra Escartín et al. (2017) drew attention to a list of potential negative effects and ethical issues concerning shared tasks in NLP. In this section, we take this list as a starting point and examine each problem with respect to the main goal of shared tasks—to advance research in the field. Some issues were regarded as potential concerns rather than definite problems, because it is unclear how large their actual impact is. We believe that some of these concerns needed to be quantified in order to be properly assessed.

To this end, we reviewed about 100 recent shared tasks from various campaigns (SemEval, EVALITA, CoNLL, WMT, and CLEF) between 2014 and 2016. We focused on several aspects, such as participation of companies, participation of organizers, closed versus open tracks, and the submission of papers by participants. Note that this is not an exhaustive overview of all shared tasks in NLP, but rather an arbitrary sample to investigate general trends in recent times. We use information drawn from this annotation exercise for assessing some of the problems we report in the following sections. The figures that are relevant for the discussion are reported in Table 1. The annotated spreadsheets used to collect this information are publicly available, together with some basic statistics and additional explanations.[2]

2.1 Life Is Not Always Fair (and Neither Is Research)

Some potential concerns, although being ethically relevant, are not necessarily a problem in terms of research advancement, and fixing them directly should not be a priority. Here, we assess issues raised by Parra Escartín et al. (2017) that we believe fall into this category.

[1] http://www.ethicsinnlp.org.

[2] https://bitbucket.org/robvanderg/sharing.

Table 1
Statistics on participation of organizers and companies in our sample of shared tasks. The "Avg. participants" column shows the average number of participating teams. We also show the average normalized rank, where the normalized rank of a team is given by $(\textit{ranking of team} - 1)/(\textit{number of teams} - 1)$. The column "Paper" indicates the percentage of participating teams that have submitted a system description.

	# Teams	Avg. normalized rank	Avg. participants	% Paper
No company	732	0.52	12.1	89.51
Some company	85	0.22	8.6	95.29
Only company	93	0.23	7.5	96.77
Organizer	56	0.08	3.5	98.21

Potential Conflicts of Interest. Parra Escartín et al. (2017) state that participation of organizers or annotators in their own shared task raises questions about inequality among participants, as organizers have earlier access to the data than the regular participants. In our survey, we found that in 5.8% of shared tasks, organizers did indeed participate. However, we also observe that this happens in connection with few participants (average 3.5 compared with 12.1, see Table 1), thus typically smaller shared tasks. This indicates that organizers' participation is more common in small, specialized tasks. The low number of participants can also explain why organizers perform better on average compared with non-organizers (see average normalized rank in Table 1).

Unequal Playing Field. An unequal playing field mainly reflects the starting point that the different teams have. Parra Escartín et al. (2017) report on the issue of differences in processing power. An extreme example of this issue is the submission of Durrani et al. (2013) at WMT13, in which they reached the highest scores because they were able to boost the BLEU score by approximately 0.8% by making use of 1TB RAM, which was probably unavailable to the other teams at that time. Computational resources are not the only reason for an unequal playing field, though. There are many other causes that could lead to an unequal playing field—for example, some teams might have access to more proprietary data, proprietary software, or research equipment.

2.2 When Competition Does Not Align with Advancing Research

The competitive nature of shared tasks can be fun, and stimulating for a variety of reasons (visibility, grant applications, beating state of the art, etc.). Such reasons might not necessarily be positively correlated with advancing the field, though. Here we discuss issues also raised by Parra Escartín et al. (2017) that we believe fall into this category.

Secretiveness. As a result of the competitive nature of shared tasks, it can be desirable for participating teams to keep their "secret sauce" private, as this could mean an advantage for a re-run of the same task, or a shared task on a related problem. As a possible effect of secretiveness, Parra Escartín et al. (2017) also mention "Unconscious overlooking of ethical concerns," actually referring to an unacceptable level of vagueness in papers. In other words, participants may unintentionally describe their systems in an abstract and vague way due to a previously established practice in systems' descriptions.

Lack of Description of Negative Results. Given that negative results are informative, their under-representation in shared tasks is a concern. A lack of knowledge about negative results might lead to a research redundancy, which is clearly undesirable. The issue of under-represented negative results is a global concern for the entire field, and for science in general. However, shared tasks provide an excellent opportunity for publishing negative results, as the acceptance of papers for publication does not particularly favor positive results.

Redundancy and Replicability in the Field. Parra Escartín et al. (2017) raise issues concerning two types of redundancy, (a) when optimal parameter settings of a previous shared task do not carry over to the new version of the task, therefore it is not clear what is learned; and (b) when algorithms are reimplemented for replicability purposes.

Regarding (a), we think that differences in used parameter settings are not actually a bad thing; we learn from this that we overfit on the previous task, or that we need to adapt our systems to another data set or domain.

Regarding (b), this is a real problem because starting from scratch to reimplement existing systems is unnecessarily time-consuming. In addition, it would always be desirable to be able to directly reproduce the same results of the same model for the same task (Pedersen 2008; Fokkens et al. 2013).

Withdrawal from Competition. Participants may withdraw from a shared task if their ranking in the competition can negatively affect their reputation and/or future funding. For example, Parra Escartín et al. (2017) suggest that companies might prefer to withdraw from the competition if they are not highly ranked, to avoid blemishing their reputation. This is something that we could not quantify in our survey, as in case of withdrawal there would be no evidence of participation in reports. There are two aspects, though, that we can quantify. The first aspect is the number of teams that do not publish their system's description, which amounts to approximately 9%, and could indeed be related to withdrawals. However, exactly because the paper is missing, information on why a team withdrew is not available. The second aspect is the total number of industry participants, which in our sample amounts to 20% ("Some company" and "Only company" in Table 1). Thus, although there is not much that can be done about withdrawal—and this might not be a problem anyway—we believe that, considering the substantial presence and interest of industries so far, their participation should be accommodated.

Potentially Gaming the System. Shared tasks are usually bound to data sets and evaluation metrics. This could lead to competition-oriented participants focusing more on tuning their systems on a given data set and metrics rather than finding a scientifically sound and scalable method for solving the problem. This can, in turn, result in an "unfair" ranking or a misleading relation between a methodology and its value with respect to the research problem. A potential negative outcome of the latter is a scenario where "optimal" methods of a shared task do not carry over to related shared tasks. These problems become more severe when system gaming is combined with a secretive attitude. While tackling this issue, we should take into account that participants might be less eager to write about ad hoc solutions, for example tuning pre-processing components or tailoring a system too closely to specifics of the annotation.

3. The Future of Shared Tasks: Sharing Data, Systems, and Failures

Our proposal for future shared tasks is not revolutionary. It simply revolves around the key aspects of *sharing*, not only resources but also experiences, including negative ones. Specifically, with research progress in mind, we believe sharing should be encouraged and even partially enforced. We therefore suggest an explicit setting for shared tasks in NLP, and reflect on the issue of what organizers could do in order to maximize sharing of information regarding participating systems. We also show how such a simple strategy can help to overcome the problems raised that can hinder research advancement.

3.1 Public versus Proprietary Track

One of the challenges faced by shared tasks is to ensure a level playing field permitting a transparent comparison of the merits of different *methods*. The problem is that system A might come out on top of system B not because its method is superior, but because, for example, it was trained on more data. This would favor teams with access to more resources, like companies with large quantities of proprietary in-house data.

Traditionally, this problem has been mitigated by establishing "closed tracks." In closed tracks, participating teams are not allowed to use any training data other than that provided by the shared task organizers. The rationale behind this is that if all systems use exactly the same data, the playing field is equal, and the competition results will show the strengths of the different *methods*. In order to study the effect of additional training data, many shared tasks have a separate competition, the so-called "open track."

However, this open–closed division is increasingly impractical and ineffective. The main problem is that it is only concerned with training *data*, whereas the performance of systems can crucially depend on other resources. Examples are external components with pre-trained models, such as part-of-speech taggers and dependency parsers, auxiliary data-derived resources like word embeddings, and other influential factors like the availability of computational resources. Because such models are almost always derived from external language data, it is unclear where to draw the line between closed and open. Should such data be disallowed or not? If not, teams still do not really participate on an equal footing.

From a research perspective, banning external models is completely impractical and nonsensical, as most state-of-the-art systems now depend on them. Likewise, trying to force all teams to use the same set of external models, and no other, would place a heavy burden on both organizers and participants. Moreover, restricting the resources participants can use is questionable, because, for research to progress quickly, teams should use the best resources available, or the resources best fitting their system.

It is therefore unsurprising that the use of closed tracks has declined in shared tasks in general in the last few years, as we have observed during our review of shared tasks. However, the original problem of unequal playing field, and thus a bias in favor of teams with ample resources, remains.

We propose, then, to rethink the problem, not in terms of equal training data, but in terms of equal opportunities. This is closely connected to the wider issue of reproducibility and replicability: Like all published research, shared task results should ideally be fully reproducible by anyone (Pedersen 2008; Fokkens et al. 2013).[3] Moreover, it should be easy to build on others' work to try out new variations of a method, without having to reimplement things from scratch. To ensure this, it is desirable that everything needed to reproduce experimental results is publicly and freely available, including code, data, pre-trained models, and so on. Interestingly, at the CoNLL-2013 shared task a similar step was taken, but only in terms of pre-condition: "While all teams in the shared task use the NUCLE corpus, they are also allowed to use additional external resources (both corpora and tools) so long as they are publicly available and not proprietary" (Ng et al. 2013). We would like to take this a step further, by enforcing the sharing of whatever resource teams might choose to use, so as to favor the injection of new resources in the field.

Applying this principle to shared tasks in practice, we propose making the primary competition a "public track," where participants can use any code, data, and pre-trained models they want, as long as others can then freely obtain them. In other words: All resources used to participate in the shared task should be subsequently shared with the community. Although this does not ensure equal access to resources for the current edition, it will still ensure a progressively more equal footing for the future. We believe this is the crucial step to move the field forward, as everyone will have access to the resources used in state-of-the-art systems. To keep participation possible for teams who cannot or will not make all resources available, a secondary, "proprietary track" can be established.

3.2 Reporting Negative Results

Ranking of systems forms a large part of the appeal of shared tasks. However, rankings should not be overemphasized and are far from being the final goal of shared tasks. Research is supposed to teach us about the merits and characteristics of methods, including insights of what does not work, rather than about which team built the system that performed best on the test data.

Negative results are very informative for future developments. Although publishing negative results is difficult, shared tasks do provide the ideal context for disclosing and explaining low performance methods and choices. We believe that shared task organizers should explicitly and strongly solicit the inclusion of *what did not work* in the reports written by participating teams. This could be even solicited via an online form that participants submit after the evaluation phase, where they comment on what worked well (as commonly done), but also provides a separate section to explain what did not work. This information could in turn be valuable data for organizers when compiling the overview report. Moreover, a clear explanation of what did not work, in connection with availability of code, would help to better understand whether something does not work as an idea or because of a specific implementation.

More generally, organizers should encourage—and to some point ensure through the reviewing process—that all participants provide exhaustive reports, potentially including ablation/addition tests, so as to have a picture as comprehensive as possible. Because participating in shared tasks directly implies getting a paper accepted for

[3] In some shared tasks, systems must be submitted through an evaluation platform. This is the case for all the tasks organized within the PAN framework (`http://pan.webis.de`), where teams must use the TIRA platform (`http://www.tira.io/`) to run their systems on test data (Potthast et al. 2014). This practice surely enables better reproducibility of results, but does not ensure sharing.

publication, not everyone describes their system to the satisfaction of external reviewers. This should change, and acceptance should be conditional on clarity and exhaustiveness.

4. Take-Home Message

We stated that the goal of advancing research should be prioritized over competition. This is especially the case when focusing on the competition aspect would encourage undesired practices like secretiveness and gaming the system. We suggest a simple solution based on the principle of maximizing resource- and information-sharing. As a byproduct, some problematic competition-related issues will be overcome, too. Some outstanding ethical issues cannot be solved, as they are intrinsic in human nature and cannot be controlled for by means of specific guidelines.

Introducing proprietary and public tracks will stimulate participants to release their systems and resources. This will directly reduce Secretiveness and the issue of Redundancy and Replicability in the Field. It will also partially address the Unequal Playing Field problem, at least in the long run: Even if at the same competition different teams will have access to different resources, all resources will be available to everyone for the next round. Moreover, being able to access and run systems on different data sets will uncover limitations that might have been due to tailoring systems to the specifics of a given shared task (Potential Gaming the System). The presence of a proprietary track still allows for industrial participation (see Withdrawal from Competition in Section 2.2), where distribution of resources might not be as easy as for other teams.

Encouraging participants to write comprehensive reports that include negative results will be a valid instrument towards advancing research, at the same time solving some outstanding problems. The reviewers should probably spend extra time in assessing the single reports and accept them conditionally on clarity requirements, but we believe this is worth the effort. Indeed, enforcing that systems are described properly will ensure transparency and increase the chances of reproducibility.

Lastly, there are some issues left unsolved. We believe these are issues that cannot or need not be solved. Withdrawal of teams cannot be controlled for. Surely, if reporting of negative results is encouraged and becomes common practice, it is possible that fewer teams will choose to leave the competition. Emphasis on progress and sharing rather than winning will also help. The conflict of interest is not relevant in our view. We do believe that organizers should be allowed to participate, and this is especially true for shared tasks that might attract a small number of participating teams due to the nature of the task. As long as these cases are explicitly reported in both the overview paper and the system descriptions,[4] the interpretation of results and ranking is easily left to the readers.

Will eliminating closed tracks also eliminate an equal playing field? And will it make it harder to truly compare different methods over the exact same data? We do not think so. Equality will be increasingly guaranteed by resource sharing, to the extent that it is possible, as inequality is intrinsically part of the game. As we argued in Section 3, comparison of methods has not been transparent in closed tracks anyway. Additionally, limiting the use of resources is not progress-friendly. More clarity in reports and release of systems will make it possible for the larger community to assess which methods work and which do not, and to improve progressively on the state of the art.

[4] This was not always the case in the sample we surveyed, but should be encouraged.

We hope that our reflections and suggestions will yield further discussion on shared tasks, to make them ever more useful for driving progress in NLP research. Each subcommunity and each task will eventually continue to have their own settings that the organizers will deem most appropriate. However, we do believe that requiring participants to release their contributions, in terms of systems, resources, and disclosure of successes and failures, should be a common trait to all.

Acknowledgments

These thoughts stem from long conversations we had, together and individually, on various occasions with many colleagues. It is impossible to thank everyone for their input. However, we want to mention a few who have mostly contributed to turning random reflections into better shaped ideas. The discussion we had at one of the weekly meetings of the Computational Linguistics group at the University of Groningen was the actual trigger for writing this piece—we are grateful to everyone present, and in particular to Antonio Toral and Rik van Noord for their sharp and valuable criticism. Viviana Patti has repeatedly and substantially contributed to the reflections on what does not quite work with the open versus closed track setting as it currently stands. We thank Johannes Bjerva, Johan Bos, James Curran, Gertjan van Noord, and Carla Parra Escartín for commenting on earlier drafts of this article. We are also grateful to Paola Merlo for her encouragement.

References

Durrani, Nadir, Barry Haddow, Kenneth Heafield, and Philipp Koehn. 2013. Edinburgh's Machine Translation Systems for European Language Pairs. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 114–121, Sofia.

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 Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia.

Ng, Tou Hwee, Mei Siew Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia.

Paroubek, Patrick, Stéphane Chaudiron, and Lynette Hirschman. 2007. Principles of evaluation in natural language processing. *Traitement Automatique des Langues*, 48(1):7–31.

Parra Escartín, Carla, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia.

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

Potthast, Martin, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2014. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In *Information Access Evaluation Meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Valencia.