Should we pursue SOTA in Computational Creativity?

Anna Jordanous
School of Computing
University of Kent
Canterbury, Kent, UK
a.k.jordanous@kent.ac.uk

Abstract

Should we pursue a *state-of-the-art* in Computational Creativity? The activity of 'SOTA-chasing', or working towards beating performance standards achieved by the current state of the art, is typical in many research disciplines relevant to computational creativity such as Machine Learning or Natural Language Generation (SOTA). Computational Creativity (CC) research does not typically engage with SOTA-type benchmarks. Consequently, it becomes harder to objectively identify high-performing systems in a creative domain (area of creative application), despite our research efforts building significant bodies of work in several domains. This paper critically engages with the use of SOTA in other related disciplines and explores the idea of working with SOTA-based evaluation in CC. The paper offers recommendations for (careful) use of SOTA to invigorate and direct CC progress.

Introduction

Should we pursue a *state-of-the-art* in Computational Creativity? In many AI disciplines related to computational creativity, typical research practice includes some evaluation experiments to compare research results to a ground truth set of results derived from some comparable benchmark or leading system in the same research area, referred to as the current state-of-the-art (SOTA). In Computational Creativity, for various reasons, the idea of a SOTA has frequently been dismissed as irrelevant and/or unachievable, despite our research efforts building significant bodies of work in several domains (areas of creative application). The consequence is that it becomes harder to identify which are the leading systems in a creative domain, in terms of inspiration or in terms of representing the bar that has been set for achievements and knowledge advances in computational approaches to creativity in these domains.

SOTA and its use in AI research

SOTA stands for State Of The Art, and refers to some leading benchmark or system for a particular task. In many AI disciplines relevant to computational creativity, such as Machine Learning or Natural Language Generation, it is typical to perform at least some evaluation in comparison to a ground truth baseline or set of results derived from the current state-of-the-art (SOTA) for that research task. This has become standard practice, to the extent that the acronym SOTA has become a recognised noun in AI research vocabulary. SOTA is typically measured objectively, either numerically or as a percentage, via metrics that have come to be recognised as appropriate for that task. Common metrics include accuracy and specificity, statistical tests, or F-scores (a combinatory measure of precision and recall).

What has also become standard practice in such disciplines is the activity of 'SOTA-chasing', or trying to better the performance of the current state of the art. This is typically encouraged. The guidelines for the International Joint Conference in Artificial Intelligence (IJCAI), a leading AI conference,[1] refer its reviewers to guidance (Blockeel and Davis 2022) that asks reviewers to evaluate experiments in a paper based on various criteria such as "'Are competitors SOTA? Are all competitors chosen? If not, how have they been selected? Are the conclusions aligned with this selection? ... this information is relevant for assessing how convincing the experimental results are" (Blockeel and Davis 2022, slide 41).

Historical perceptions of SOTA in CC

Computational Creativity is not a discipline where we tend to record or measure any state-of-the-art. Within the field, objective evaluation metrics based on the product of a creative system such as Ritchie's empirical criteria (Ritchie 2007), once quite popular, are now not used very often. Such objective evaluation was criticised for only evaluating the product of creative systems, ignoring the process by which they operated (Colton 2008), and other of the Four Ps of creativity (Jordanous 2016) (Producer, Press, Product, Process). Ritchie's criteria also required some agreement on domain-appropriate choices of threshold values and parameters for the criteria. However we have seen Ritchie's criteria deployed with some success for comparative evaluation in the areas of narrative generation (Pereira et al. 2005) and music improvisation (Jordanous 2012).

Other generic evaluation metrics or frameworks exist such as FACE (Pease and Colton 2011) and the Creative Tripod (Colton 2008), or domain-specific evaluation metrics such as O'Donoghue's statistical tests for analogies (O'Donoghue

[1] https://ijcai-22.org/faqs/

2007). These tend to be implemented using subjective judgements, difficult to replicate consistently for comparison over time due to possible variability in human opinion.

SOTA: Meaningless numbers?

If we did try to deploy some sort of objective metric for evaluation in CC, what would the measurements actually represent? Wouldn't numeric measurements or percentages be meaningless? Not necessarily. Objective metrics have been proposed that could be used for comparative evaluation against some established baseline, such as the work by Bossou and Ackerman (2021), or the (to-date, unused) IDEA model (Pease and Colton 2011) and previous tests proposed by Pease, Winterstein, and Colton (2001). It is also not impossible to consider ways in which methods such as FACE and the Creative Tripod could be operationalised in objective metrics. The SPECS evaluation methodology (Jordanous 2012) also opens up ways for evaluative tests to be defined relative to a definition of creativity, which could be defined objectively. We have seen specific uses of evaluation metrics defined for particular domains (areas of creativity), such as the use of scores for story plots (Pérez y Pérez 2014).

Comparative evaluation as a blunt tool in CC?

What does it mean to measure or compare one system against each other? It seems unrealistic to pursue the devising of universal SOTA benchmarks that might cover all different types of creative systems. But that should not stop us in our tracks. Fields such as Machine Learning use SOTA benchmarks to compare applications or algorithms that work on roughly the same task, that can be directly compared.

Do we have enough effort in particular applications of creativity to have a meaningful domain-specific SOTA benchmark for that area? While we have seen arguments for (and evidence of) comparative evaluation being useful to measure progress (Jordanous 2012, e.g.), a common feeling in earlier days in CC was that it does not make sense to evaluate systems against each other, as we did not have enough comparable systems to establish a state of the art. CC has now reached a stage, however, where there are various application domains that are well represented in terms of different systems (Loughran and O'Neill 2017).

A more subjective objection might be that it feels to some extent inappropriate to have a system identified as best-performing in a specific domain of creativity, due to the wide variety of ways in which creative systems can excel even if performing comparable tasks. (We should acknowledge that this has not stopped the existence of human-equivalent competitions of the 'best' artist, or story-teller, or idea generator, for example, nor the monetary valuing of creative outputs.)

But without recognising the achievements of some systems as superior to others, how can we hope to learn from the systems that do outperform others? Let us consider the potential benefits of some kind of SOTA-based evaluation.

Potential benefits of SOTA evaluation

If we could use SOTA-based evaluation in CC, would the field benefit? In other words, if we could establish metrics that captured a state-of-the-art baseline in various domains that are well-covered by Computational Creativity research, such as narrative generation, visual art generation, or music composition, then what would we gain from testing new systems in those domains against the current state-of-the-art? Learning from other disciplines that use SOTA, we could have tangible ways to measure progress in particular research domains (Lewis and Crews 1985). This might help computational creativity research venues to establish greater credibility within more general AI conferences such as IJCAI, ECAI, AAAI and so on, where our typical papers may not currently be seen as containing enough rigour due to lack of comparative experiments against a SOTA. Perhaps more importantly, if we could establish SOTA for a particular CC domain, then this would be transferable to those working outside of the direct CC community. Looking at conferences in the remit of computational creativity such as ISMIR (music) or ACL (language), it is still possible to have papers accepted with 'hand-wavy' justifications of a system being considered creative with little or no rigorous evaluation of that claim of creativity; because (within my own subjective experience) there is little adoption of CC's creativity evaluation metrics outside of the CC field itself.

Does 'SOTA-chasing' give us a clearer idea of the best current systems in a particular area? And when a new system represents a significant advance? After all, our current ways of identifying the current state of the art are subjective, hence vulnerable to misinterpretation and bias.

There is of course significant pressure to get appropriate metrics of strong performances in a creative domain. Pursuing a SOTA benchmark for a domain could help us establish objective metrics for evaluation, available for reuse to compare systems (typically considered good practice in terms of establishing how systems represent advances in knowledge).

Potential risks of SOTA evaluation

Use of SOTA evaluation in AI/ML areas is common, and accompanying this is the risk of getting only minor incremental advances - where papers could be considered ready to publish if they advance SOTA by a minuscule percentage. At the other end of this extreme, we are a field which typically encourages innovation in method and approach even if there is not a tangible effect on results; we do not want to be in the situation where a system that does not beat SOTA becomes almost unpublishable.

'SOTA-chasing' as a research activity has been criticised by some (Church and Kordoni 2022; Koch et al. 2021). One criticism of particular relevance to CC is the question of what approach to take if we do not have a direct or obvious-fit metric to use. There is no one 'test for creativity'. In this circumstance, we can examine what another similar field does. Thanks to the likes of the GPT-* transformer systems et al, deep learning-based text generation has seen phenomenal progress over the past few years. Typically, such systems need to evaluate output at scale, with large data output to evaluate that needs automated metrics. Lacking a specific automatable metric for evaluating generated text (a problem familiar to those working with creative language generation), it is common to see the machine translation metric

BLEU used as a *proxy* for evaluating the success of a system in learning from some input data to generate some output data. In other words, such a metric is considered to be an approximate evaluation of success: 'good enough' to be adopted in order to facilitate progress.

What happens if we use the wrong metrics, or fail to evolve or adapt our metrics over time as needed? The reliability of experimental research depends heavily on how research findings are derived from scientific experiments (Ioannidis 2005). Does research go in the wrong direction? Taking our example of transformers research, only time will tell, but the phenomenal progress over the past few years seems to suggest that the adoption of a 'good enough' proxy metric has been helpful in allowing research to progress. In such situations, community self-awareness of the metric's status as a proxy, not a direct measurement, is critical.

Recommendations for use of SOTA in CC

Would SOTA chasing be enough to replace other evaluation? No, probably not, particularly as this would reverse progress in creativity evaluation and risk us forgetting what we have learned about evaluating our creative systems (see the *Historical Perceptions* discussion above). *But it could complement what we are already doing.*

We should acknowledge that even in disciplines where SOTA-based evaluation has come to be typical, it is not mandatory for research success and such research communities do not always advocate for experiments referencing and comparing to SOTA. Although, as remarked above, the IJCAI conference refers reviewers to recommendations (Blockeel and Davis 2022) to check if experiments compare a piece of work against SOTA, the same guidance also states what to do if you are reviewing a paper where there is:

> " "No experimental comparison to SOTA". Ask yourself: is it needed?
>
> - In 95% of cases: yes. But be aware of that 5%.
> - e.g.: theoretically very innovative work, novel insights, ... may be valuable even if an implementation or experimental comparison is not possible at this time"

(Blockeel and Davis 2022, slide 39)

The *Historical perceptions* section above reflects on how we could implement SOTA metrics in ways which do not focus just on measurable aspects of creative output, but which measure process and other Four P perspectives. In some contexts, a SOTA benchmark would be establishable with current metrics (fine-tuned towards some objective metric, as discussed above). In fact, it could be argued that this has already happened in the context of poetry evaluation (Pereira et al. 2005). We could delve into the statistical and empirical measurements and tests common in AI and data science, and see what could be used, as for example in O'Donoghue (2007). There are other measures of subjective concepts that we could learn from and perhaps re-appropriate for SOTA metrics, for example, Seth's measure of autonomy and emergency (Seth 2010).

Proposal: CC-eval competition

In research areas such as music Informatics, NLP, and multimedia computing, as well as (even closer to home for CC) procedural content generation, research progress is aided by evaluation against benchmarks, as part of regular (typically annual) competitions. See for example:

- MIRAX (music)
  `https://www.music-ir.org/mirex/`
- SemEval (NLP)
  `https://semeval.github.io/`
- MediaEval.org (multimedia computing)
  `http://www.multimediaeval.org/`
- GDMC - Generative Design in Minecraft (PCG)
  `http://gendesignmc.engineering.nyu.edu/`
  - (and GDMC's Chronicle for games narratives)

Does CC need a CC-eval competition like MIREX, SemEval, and so on? We have in the past seen curated exhibitions in past ICCCs, so we have an established vehicle within which to host such an event each year. And let it be remembered that we do already do competitions, or at least those of us working in PCG do. The GDMC competition has seen considerable growth in the few years it has been operating, acting as a high visibility route into established and well-defined PCG challenges. Treating GDMC as a test case, it's important to recognise that the use of metrics based on human judgement requires a lot of effort on the part of judges. This has led to exciting work with GDMC organisers exploring automatable metrics (Hervé and Salge 2021)

How could a CC-eval competition work? This could follow a MIREX-like model of proposed tasks each year, many of which may re-occur from year to year. In this model, the task proposers also propose evaluation metrics that are applied to all entries (and any 'inspiring set'/training data).

Such a competition could provoke interest in pre-defined tasks (as GDMC and SemEval/MediaEval/MIREX do), with potential benefits of attracting new researchers and also keeping established researchers engaged (and challenged by the 'new kids on the block'!) Such competitions have seen their tasks form the basis of student projects at undergraduate level and above. They have been useful for community spirit building and the establishment of GroundTruth metrics by those working directly in a creative domain who feel confident enough to propose and run the task that year. Metrics could be examined and used every year that a task runs.

This proposal comes with downsides, of course. We would need to tackle many challenges outlined in this paper, particularly if proposing a task. Initial task metrics would require some very careful thinking, ideally crowdsourcing via experts in that CC domain. For subjective evaluation metrics, could we get enough commitment from judges? MIREX have in the past struggled with this, for example. There would be considerable obstacles in terms of set-up effort, time commitment, organisational infrastructure and reliance on volunteers, at a time when many of us are exhausted and burnt-out from pandemic related uncertainties and workloads. But perhaps this would help us come together to reinvigorate that part of our community spirit that is so hard to replicate if not meeting every year in person, as well as create an exciting entry point for newcomers?

Conclusions

The field of Computational Creativity has thus far resisted the idea of establishing and bettering a current state-of-the-art target for specific domains. SOTA-chasing has become the norm in various sub-fields of AI such as Machine Learning (ML) or Natural language Generation (NLG). As commented above, recent advances in NLG provide an example of the remarkable progress that can be facilitated through using SOTA benchmarks for targeted improvement, even when metrics are not as clearly identifiable as in tasks which can be measured using statistical or information-theoretic measures.

My argument in this paper is that meeting or beating SOTA in CC is not the requirement it is billed to be in ML, and it also is not the devil it could sometimes be perceived to be in CC. I suggest CC research has reached a point of maturity where we can start doing it, to help us track progress in each creative domain that we have built up a body of work in. This will help build the field, as long as we can learn from those in related disciplines and avoid weakening our research due to falling into the traps identified by Goodhart's law - "when a measure becomes a target, it ceases to be a good measure" (Oxford Reference retrieved May 2022).[2]

There are many pitfalls to be aware of. What I propose here should not replace more substantial evaluation, but could complement it. Pursuit of SOTA metrics could help us in the pursuit of evaluation metrics, as well as adding a new way to target and track progress and even help build our community further. I posed a possible route forward of a CC-Eval competition, as a *Grand Challenge* for CC, inspired by the likes of MIREX and SemEval (but I should stress this is one of many possible routes forward).

We should acknowledge that metrics for measuring SOTA in a creative domain may need to change over time, to avoid the criticism that credibility of a scientific field of research is weakened by lack of flexibility for that field to self-correct (Ioannidis 2012). As one reviewer of this paper commented, we also need to be familiar the meanings and intentions behind the metrics we use, to critically appreciate the levels of meaningfulness and informativeness of results.

Our research community (and domain sub-communities) contain enough domain expertise to recognise and collectively establish the most appropriate metrics for a creative application area. As a community, we have a history of engaging enthusiastically with self-reflection and self-correction (see for example the paper types in the Short Paper call for this conference). We also have a history of considering evaluation of creativity deeply, including metrics for meta-evaluation that we could apply to our tests for SOTA benchmarks (Jordanous 2014).

What we do need, to progress this further, is for people working in specific areas of computational creativity to propose, use, evolve and convalesce onto some SOTA metrics for those areas. These metrics do not need to be perfect; we know this is pretty much impossible in many creative domains. However careful choosing of 'good-enough' metrics as a proxy for that creative area - as the text generation community have done - opens doors for tracking and furthering progress in various domains of Computational Creativity.

Author Contributions

AJ ideated and wrote the paper alone, though as acknowledged below, the paper benefitted much from discussions.

Acknowledgments

The inspiration for this paper comes out of failed attempts to explain to my colleagues at the School of Computing, University of Kent, why CC does not pursue state-of-the-art comparisons. I also thank Piotr Sawicki, Marek Grzes, Dan Brown and Max Peeperkorn for our conversations around the use of SOTA. In particular, conversations with Max Peeperkorn prompted me to finally write up these ideas for ICCC'22, after my pondering on them for some time.

References

Blockeel, H., and Davis, J. 2022. A brief introduction to reviewing. https://dtai.cs.kuleuven.be/introduction-to-reviewing-for-ijcai.pdf , last retrieved 22 April 2022.

Bossou, K., and Ackerman, M. 2021. Should machines evaluate us? opportunities and challenges. In *Proceedings of the 12th International Conference on Computational Creativity, Mexico City, Mexico.*

Church, K. W., and Kordoni, V. 2022. Emerging Trends: SOTA-Chasing. *Natural Language Engineering* 28(2):249–269.

Colton, S. 2008. Creativity versus the Perception of Creativity in Computational Systems. In *Proceedings of AAAI Symposium on Creative Systems, Stanford, US*, 14–20.

Hervé, J. B., and Salge, C. 2021. Comparing PCG metrics with Human Evaluation in Minecraft Settlement Generation. *ACM International Conference Proceeding Series.*

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine* 2:e124.

Ioannidis, J. P. A. 2012. Why science is not necessarily self-correcting. *Perspectives on Psychological Science* 7(6):645–654.

Jordanous, A. 2012. A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative. *Cognitive Computation* 4(3):246–279.

Jordanous, A. 2014. Stepping back to progress forwards: Setting standards for meta-evaluation of computational creativity. In *Proceedings of the 5th International Conference on Computational Creativity, Ljubljana, Slovenia.* Ljubljana, Slovenia: ACC.

[2] The excellent comments from the anonymous reviewers, including the reference to Goodhart's law, demonstrate how CC researchers can - and do - engage very productively with this debate, even if one does not agree with the arguments I present here.

Jordanous, A. 2016. Four PPPPerspectives on computational creativity in theory and in practice. *Connection Science* 28(2):194—-216.

Koch, B.; Denton, E.; Hanna, A.; and Foster, J. G. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Online*.

Lewis, B. C., and Crews, A. E. 1985. The evolution of benchmarking as a computer performance evaluation technique. *MIS Quarterly* 9(1):7–16.

Loughran, R., and O'Neill, M. 2017. Application Domains Considered in Computational Creativity. In *Proceedings of the Eighth International Conference on Computational Creativity (ICCC'17), Atlanta, GA*.

O'Donoghue, D. P. 2007. Statistical evaluation of process-centric computational creativity. In *Proceedings of the 4th International Joint Workshop on Computational Creativity, London, UK*.

Oxford Reference. retrieved May 2022. Goodhart's law. https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095859655.

Pease, A., and Colton, S. 2011. Computational Creativity Theory: Inspirations behind the FACE and the IDEA models. In *Proceedings of the 2nd International Conference on Computational Creativity, Mexico City, Mexico*, 72–77.

Pease, A.; Winterstein, D.; and Colton, S. 2001. Evaluating Machine Creativity. In *Proceedings of Workshop Program of ICCBR-Creative Systems: Approaches to Creativity in AI and Cognitive Science, Vancouver, Canada*, 129–137.

Pereira, F. C.; Mendes, M.; Gervás, P.; and Cardoso, A. 2005. Experiments with Assessment of Creative Systems: An Application of Ritchie's Criteria. In *Proceedings of the Workshop on Computational Creativity (IJCAI 05), Edinburgh, UK*.

Pérez y Pérez, R. 2014. The Three Layers Evaluation Model for Computer-Generated Plots. In *Proceedings of the Fifth International Conference on Computational Creativity, Ljubljana, Slovenia*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Seth, A. K. 2010. Measuring autonomy and emergence via Granger causality. *Artificial Life* 16(2):179–196.