Effortful & Expert Evaluation in Developing Serious CST

Diarmuid P. O'Donoghue
Department of Computer Science
Maynooth University, Ireland
diarmuid.odonoghue@mu.ie

Introduction

This paper outlines some challenges involved in developing *creativity support tools* (CST) aimed at *serious creativity*. Such domains include: academic research, patent creation, literature-based discovery, creative ideation, *etc*. By referencing a common workflow model, we focus on the role of evaluation within the development cycle and finessing of a CST. Our focus lies in gathering expert evaluations by recognised leaders and critics whose opinions hold respect within that community. Such evaluations are typically; difficult to acquire, involves experts with very narrow field of expertise and necessitate detailed and complex evaluations. We outline an approach to evaluation that is based on pre-selected evaluators for whom personalised artefacts are created for evaluation.

Serious Creativity

Developing *creativity support tools* (CST) for *serious creativity* can often find it difficult to attract suitably experienced, leading to great difficulty in obtaining enough useful and action-able evaluations to guide development. Difficulty in accessing suitable evaluations can lead to significant impact on the development of these CST tools. This paper identifies some impediments to developing such CST and identifies some possible approaches to a range of these problems.

Pease *et al* (2019), O'Donoghue and Keane (2012) and others have highlighted the importance of scientific and mathematical disciplines to computational creativity. The Creativity Achievement Questionnaire - CAQ (Carson *et al*, 2005) identifies several domains that are likely to require effortful and expensive evaluation - including CAQ's "areas of talent".

1. Scientific discovery - scientific hypotheses, mathematical conjectures, literature-based discovery (Juršič et al, 2012)
2. Invention - patent creation
3. Architecture - including Engineering
4. Entrepreneurial Ventures - a CST for creating business plans
5. Educational support (Goel and Joyner, 2015) - to support student centred discovery
6. Journalism to create new story angles (Maiden *et al*, 2020)

Domains where some evaluations can be easily obtained include: Visual arts, Music, Creative writing, Dance, Drama, Humor and Culinary Arts. This list could include individual sport and team sport from the CAQ's additional "areas of talent" list.

However, some creations from these domains may also require effortful evaluation. For instance, if we wish to collect expert critiques for a serious artist (the CST), we might expect a similar challenge in acquiring such professional evaluations. Similarly, a CST (or even autonomous creativity system) for the culinary arts may be evaluated not by general members of the public, but by expert critics and might encounter some of these challenges.

CST Workflow Model

The envisaged development process lies in contrast to developing many CST that can rely on free and easy access to a plentiful supply of competent evaluators. Developing CST for serious creativity often requires evaluations to guide development and the difficulty of obtaining evaluations can greatly impact the development process itself. It might even be argued that evaluation plays a more prominent role developing CST than in autonomous creative system due to the co-creativity involved in multiple steps of the workflow. Some autonomous creative systems may only involve evaluations at the final phase.

Comparing the role of evaluation across CST would greatly benefit from a standard workflow model – particularly across the artistic and scientific divide(*sic*). Wallas influential model as expanded by Sadler-Smith (2015) identifies five phases:

1) Preparation	*2) Incubation*	*3) Intimation*
4) Illumination	*5) Verification*	

We use this Wallas-Sadler-Smith model for references, but much of this paper is equally relevant to other frameworks for creativity. However, this framework's relevance to some approaches is much less apparent (GAN's *etc*).

Questions: During which phase are evaluations sought? Do evaluations cover a single phase or multiple phases? Is there just one evaluation for each artefact or multiple evaluations after each phase?

Effortful, Expert Evaluation for CST

Effortful evaluation involves significant amounts of focused time by domain experts. For example, evaluating

submissions to a journal might involve several hours of effort by an extremely narrow number of researchers. Identifying and recruiting expert evaluators is a significant challenge as many are busy professionals with limited availability. So, crowd sourcing evaluations like Amazon MTurk or Figure Eight (previously CrowdFlower) may not prove successful. For instance a CST for the Grand Challenge of developing an AI win a Nobel prize in science (Kitano, 2016) would require great access to top scientists.

Gold Standard

In place of online evaluation can serious CST make use of the gold standards that have guided much progress in AI but are not generally available or accessible for creativity. Having the CST invent its own criteria encompassing utilitarian and aesthetic values requires at least some evaluations or ground truths.

Questions: Can the wide availability of publications and patents containing links to prior work serve a basis for some form of evaluation? Can the similarities and differences between an artefact and the identified "prior work" serve to create an initial model of evaluation?

Evaluation & Verification:

Does a CST evaluate all properties simultaneously or are properties evaluate separately? Can evaluations make use of established creativity evaluations such as: CAQ, CSI (Cherry and Latulipe, 2014), SPECS (Jordanous, 2012) or simpler Likert-scale ratings. When a CST is available online this affords possible AB testing to guide the development process – but generally requires a reasonably large volume of users. Other online criteria may also play some role, such as CST interaction duration, number of returning sessions. Again however, their applicability to serious CST may be limited.

SPECS Qualities

One approach that has word involved a tailored version of the SPECS (Jordanous, 2012) qualities. Firstly, an initial campaign was run to firstly validate the importance of creativity to that community. Secondly, evaluations identified those of the 14 SPECS components of creativity that were of greatest relevance to the given community. For example, "Spontaneity/Subconscious Processing" might be of greater relevance to artistic creativity, while "Generation of Results" might be associated with scientific domains. This identified a reduced set of qualities for evaluators to consider when reviewing a artefact produced by the CST. While this approach simplified the evaluators responses, it did not seem to help in attracting the expert opinions that would really aid systems development.

Recruitment with Personalised Creativity

The challenge of recruiting of top experts to spend the time and effort to provide evaluations can be a serious problem. One approach that worked best for the Dr Inventor (O'Donoghue *et al*, 2015) employed personalised creativity. This proved to be surprisingly successful in attracting reviewers for one of the top ranked conferences across the entire discipline of computer science.

Unfortunately, it seems that this approach is only applicable when the creative process can be guided towards specific types of artefacts. Bisociative (Kostler, 1964) creativity creates novel artefacts through complex interactions between two items. Carefully choosing one of those items can often have a definite focusing effect on the created artefact – and in this case we choose one of those items to match the expertise of a potential evaluator.

In the first phase, we identified experts with the potential to act as evaluators. For scientific creativity we found resources such as the lists of reviewers for conferences and journals to be very useful, focusing on the most recent 3 years of the conference series. We identified papers written by these authors in the relevant publication venue and then used these publications to drive the personalised creativity.

This approach also offers the possibility of focusing on specific disciplines by selecting evaluators from different conference series. Additionally, this mode of personalised creativity is focused on publications – which typically have multiple authors – increasing the changes that one author might evaluate each artefact - either acting independently or responding as a group.

It must be acknowledged that this selection process introduces the possibility for bias in the evaluation process, which may (in principle) be countermanded by bias detection and correction activities. However, one advantage of pre-selecting evaluators is it enables the possibility of profiling uses to a great degree. However, recruiting evaluators for more commercially sensitive domains may be more difficult - especially when the evaluations are aimed at final finessing of a CST as any evaluated artefacts may incur a degree of commercial sensitivity.

The second phase produced creative artefacts targeted at each expert evaluator. The biscociation (Koestler, 1964) creativity behind Dr Inventor (Abgaz *el al*, 2017; O'Donoghue et al, 2015) involved exploring analogical comparisons between pairs of research publications, using a recent publication by a targeted evaluator as one of these. We identified the best analogy for that paper. Evaluations showed the preferred analogies suggesting the most inferences – with the relation between these creative inferences and the evaluators "driving" publication being highlighted.

Evaluation by Free Text

Interestingly the evaluation mechanism that worked best for Dr Inventor used free text feedback for evaluation. It was natural and easy for evaluators to provide and having invested significant amounts of time in an evaluation (typically around 30 minutes per artefact), users seemed to value to completeness of such feedback – in comparison to more constrained ratings scales and other mechanisms.

Several reasons appear to be behind the success of this evaluation drive, which centred around personalised creativity. Firstly, this approach involved a significantly reduced workload – necessitating the reading of just one publication instead of two. Secondly, the artefacts bore a direct relationship to the evaluators previous work, this connection being included in the initial contact. Thirdly, the artefact should assist the evaluator's own creativity. Finally, evaluators were familiar with providing text evaluations on submission for conferences and journals.

Metrics from Free Text

As well as being analysed subjectively, free text evaluations were converted into tree types of metric. Firstly, we performed sentiment analysis and secondly, the number of words of feedback was analysed as a separate indicator of quality. The general trend indicated that more voluminous feedback suggested quality – though not necessarily agreement with the creation. In guiding development, our preference lay in worthy ideas even if they were not perceived as technically correct – these being preferred over ideas that attracted no feedback of any type.

Consistency of Evaluations

A subset of the initial emails attracted responses from the targeted group of potential evaluators. Agreement may not always be present between the acquired evaluations, with some expert evaluations totally disagreeing on the creativity evident in an artefact. Inconsistency between evaluations represent a significant problem and using average ratings from small numbers of respondents seems counter-productive.

During development we considered the best evaluations, addressing the issue of whether the CST *ever* supported creativity. Later evaluations may increase their focus on how frequently a CST does (not) amplify users' creativity.

Questions: Can we develop models of user evaluations? Can any inductive approach across a prior collection of artefacts plus evaluations be used to predict evaluations for novel artefacts? Is this even possible for creative tasks?

Black Hat (Deceptive) Creativity:

CST for serious creativity raises the possibility of "Black Hat" creativity intended for *deception*. Fake artefacts negatively impact on other artefacts around them. Fake publications often attract a lot of attention and seriously damage the reputation of all papers in that publication venue. Can evaluation address and even counteract such concerns?

Questions: How/Can we protect against Black Hat creativity? Can evaluation play any role in guarding against such misuse of CST systems?

Conclusion

Attracting a sufficient volume of high-quality and effortful evaluations is a serious challenge for developing CST for serious creativity (science, patents *etc*). Personalised creativity aimed at a community of pre-selected (potential) evaluators can serve as a mechanism to involve evaluators and aid development.

References

Abgaz, Y.; Chaudhry, E.; O'Donoghue, D.; Hurley, D.; & Zhang, J. J. 2017. Characteristics of pro-c analogies and blends between research publications. *Intl. Conf. on Computational Creativity ICCC*.

Carson, S. H.; Peterson, J. B.; and Higgins, D.M. 2005. Reliability, validity, and factor structure of the creative achievement questionnaire. *Creativity Research Journal*, 17 (1): 37-50.

Cherry, E; and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index." *ACM Transactions on Computer-Human Interaction,* 21(4): 1-25.

Goel, A.K.; and Joyner D.A. 2015. Impact of a creativity support tool on student learning about scientific discovery processes. *Intl Conf Computational Creativity ICCC*.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems. *Cognitive Computation*, 4(3): 246-279.

Juršič, M.; Bojan, C.; Tanja, T.; and Lavrač, N. 2012. Cross-domain literature mining: Finding bridging concepts with CrossBee. *Intl. Conf. on Computational Creativity ICCC*.

Kitano, H. 2016. Artificial intelligence to win the Nobel prize and beyond. *AI magazine*, 37(1): 39-49.

Koestler, A. 1964. The Act of Creation.

Maiden, N.; Zachos, K.; Brown, A.; Apostolou, D.; Holm, B.; Nyre, L.; ... & van den Beld, A. 2020. Digital creativity support for original journalism. *Communications of the ACM*, 63(8), 46-53.

O'Donoghue D.P.; Keane, M.T. 2012. A Creative Analogy Machine: Results and Challenges, *Intl. Conf. on Computational Creativity ICCC*, pp 17-24.

O'Donoghue, D.; Abgaz, Y.; Hurley, D.; and Ronzano, F. 2015. Stimulating and simulating creativity with Dr inventor. *Intl. Conf. on Computational Creativity ICCC*.

Pease, A.; Colton, S.; Warburton, C.; Nathanail, A.; Preda, I., Arnold, D.; ... and Cook, M. 2019. The Importance of Applying Computational Creativity to Scientific and Mathematical Domains. *Intl. Conf. on Computational Creativity ICCC*.

Sadler-Smith E. 2015. Wallas' Four-Stage Model of the Creative Process: More Than Meets the Eye? *Creativity Research Journal*. 27(4): 342-352.