Statistical Evaluation of Process-Centric Computational Creativity

Diarmuid P. O'Donoghue
Department of Computer Science
NUI Maynooth
Co. Kildare
Ireland
diarmuid.odonoghue@nuim.ie

Abstract

We adopt a process-centric approach to computational creativity, based on a model of people's innate ability to process analogical comparisons. A three-phase model of analogical reasoning is adapted to function as an analogy generating machine. It is supplied with two distinct knowledge-bases containing many domain descriptions, with the aim of generating novel analogies – potentially even creative ones. However, because our approach to computational creativity does not have the usual "inspiring set", evaluating its output can not be performed by comparison to this inspiring set. Our generic approach to evaluating process-centric computational creativity uses a number of nonparametric statistical techniques. After the creative artefacts are generated, human raters assess these artefacts for the qualities of creativity (*quality, novelty etc*). We describe the results of two experiments that were conducted on these two collections of domains. The analogies generated on the two collections are analysed and difference in the two result sets are assessed. We argue that true creativity can only be assessed after the creative artefacts are generated. Evaluating creativity only by reference to the inspiring set runs the risk of overlooking creative artefacts that differ from the inspiring set - and may under-estimate a model's creativity.

Keywords: Analogical Creativity, Analogy Generation, Evaluation, Nonparametric Statistics

1 Introduction

Computational creativity frequently uses an "inspiring set" of creative artefacts (music, images, poems *etc*) both to drive the model and to act as a basis for its evaluation. This *artefact-centric* approach to computational creativity contrasts with the *process-centric* approach in this paper and elsewhere (O'Donoghue, 1997; Gomes et al., 2003; Veale, 2004; O'Donoghue et al., 2006). This paper describes an approach to computational creativity that is based on people's innate ability to understand analogical comparisons. This paper builds upon computational models of the analogical reasoning process.

Analogical comparisons are often cited as a driving force behind creativity, providing new perspectives on some previously known concepts (Boden, 1992). Creativity using analogies is strongly associated with science and scientific advancement. Pierre Curie and colleagues deliberately used analogies as a technique for generating hypotheses which they later tested (Curie, 1923). Hoffman (1995) and Brown (2003) detail the role that analogies played in many recorded scientific breakthroughs. Dunbar (2001) and Dunbar and Blanchette (2001) note that experts display a great ability to generate and use novel analogies when dealing with situations that arise in their normal work environment (this contrasts with the rare use of analogies by non-experts when presented with tester-determined analogy problems). Koestler (1964) was also among those who account for creativity as the juxtaposition of two very different sets of ideas.

In essence, an analogy is a comparison between two concepts (*source* and *target*) (Gentner, 1983), such that the source highlights particular aspects of the target and suggests new inferences about that target. Every analogical comparison has two effects. Firstly, it highlights an existing non-obvious similarity between two concepts. Secondly, it then extends this similarity by transferring information from one concept to the other, adding new information to the target. In creative analogies (Boden, 1992; Eysenck and Keane, 1995) a strange source domain conjures up a revolutionary new conceptualisation of the target, suggesting inferences that explain some previously unexplained or unnoticed phenomena.

Computational modelling of the analogy process has focused primarily on the central mapping phase (Keane et al., 1994) ; see (French, 2002) for a review). Surprisingly, only a few models have been developed of the previous retrieval phase (Thagard et al., 1990; Forbus et al., 1995; Plate, 1998; O'Donoghue and Crean, 2002; Gomes et al., 2003) or of the subsequent validation phase (Falkenhainer et al., 1989). However, no combined *retrieval–mapping–validation* model has been described and evaluated. In this paper we investigate a three-phase model and evaluate its potential for finding and assess-

4th International Joint Workshop on Computational Creativity (IJWCC), pp 31-38, Goldsmiths, University of London, 17-19 June, 2007.

ing novel analogies - potentially even identifying creative ones.

We should not be overly proscriptive in how we assess computational creativity, unpredictability being a quality that is often associated with creativity. Any creativity model that is assessed solely by comparison to an inspiring set may inadvertently overlook outputs that are considered creative when assessed independently of that inspiring set. True creativity can only be assessed *post hoc*.

The remainder of this paper is structured as follows. First we review computational creativity, distinguishing between the traditional "inspiring set" approach and our "process-centric" approach. We then describe our computational model for generating and evaluating analogical comparisons. We describe the problem of assessing our model in the absence of an inspiring set. We describe a number of statistical techniques that serve to evaluate process-centric creativity models. We then present and analyse the analogies generated by our model before drawing some final conclusions.

2 Computational Creativity

Ritchie (2001) describes and formalises the typical process by which most computational creativity programs are constructed. The process starts with basic items which are items of the type to be produced by the program (poems, music, images *etc*). A subset of these items is selected, taking into account the ratings and values associate with the basic items - creating the *inspiring set*. Following this, the program is constructed and executed for a range of parameters. We characterise this approach to computational creativity as *artefact-centric* creativity.

2.1 Process-Centric Creativity: Beyond the Inspiring Set

Our approach differs from standard computational creativity in a number of ways. First, we start with a computational model of people's ability to reason analogically. Our model is based on many years of focused work on the analogy process by many cognitive scientists. Our aim is not only to generate creative artefacts (analogies), but to do so in a cognitively plausible manner. We characterise this approach to computational creativity as *process-centric* creativity.

We reject any suggestion that producing creative analogies is somehow driven by an "inspiring set" of previous creative analogies. We do not wish to produce analogies that are similar to existing analogies. What we are searching for is analogies that produce the same effect, of explaining or highlight some facts. As noted Aristotle's Poetics[1] (Chapter 22) analogy is *"one thing that cannot be learnt from others"*.

A second difference with standard computational creativity is that we wish to produce an unconstrained model of creativity. We may expect our model to (re)generate a few well-known creative analogies (like Rutherford's *solar-system:atom* analogy), but we do not consider identifying these as *true* examples of creativity. That is because these are well studied analogies and they have been described in the literature in such a way as to maximise the similarity between the two domains. This has two implications. Firstly the expert's intricate knowledge is translated into a greatly simplified format, where the same relations are used to describe the source and target domains. Retrieving the simplified source can use identical token matching (used by MAC/FAC (Forbus et al., 1995)), but this would prove far less effective on the *original* problem as it was then understood. Even the semantic similarity metric used by ARCS (Thagard et al., 1990) may not provide a cue to retrieval. Secondly, the topology of the simplified domains are (generally) also identical, allowing the graph structure of the domains to play a significant role in retrieval (Plate, 1998; O'Donoghue and Crean, 2002; Gomes et al., 2003). Thus, identifying these analogies indirectly makes use of their original discovery - and does not require the same creative insight that is associated with Rutherford and others.

For example, O'Donoghue (1997) describes three successive problems that Kekule must have overcome in re-structuring the carbon-chain analogy into the famous carbon-ring analogy. Attempts were made to describe the domains in a manner more like the common understanding *before* Rutherford's famous insight - using the then dominant "plum-pudding" interpretation of the atom. However, the success of these attempts varied widely depending upon two factors. Firstly, the topological similarity of the resulting domain descriptions as the CWSG inference algorithm (Holyoak et al., 1994) generates inferences as a form of graph-completion process. Secondly, identicality or semantic similarity between the relations used to describe the source and target domains greatly influenced the likelihood of the correct analogy being drawn.

A creativity model should identify any additional creative analogies that might arise - ones that were not expected to be found. Thus, if one source domain offered some novel and useful inferences about some target domain, this *p-creative* (Boden, 1992) comparison should also be identified as creative.

A third difference with standard computational creativity arises from the fact that we do not begin with an inspiring set, as generated artefacts (the analogies) can not be evaluated by comparison with that inspiring set. Any process-centric model of creativity must be capable of differentiating between creative and non-creative artefacts, testing artefacts for the qualities associated with creativity; *novelty, quality etc* (Ritchie, 2001). However, this still leaves us with the task of assessing the goodness of the generated artefacts. Were any of the generated artefacts considered creative by humans?

Much of the remainder of this paper concerns this third point, evaluating the output of process-centric models of creativity. In the next section we examine our creative analogy model before turning our attention to analysing the analogies that were created.

[1] Aristotle made this statement about metaphor which is very similar to analogy, both being centred on a core mapping phase.

3 A Creative Analogy Machine

Wallas (1926) proposed a five-phase model of the creative process, composed of the phases: *preparation, incubation, intimation, illumination* and *verification*. This phase-model of creativity bears a striking similarity to many phase-models of the analogy process. Keane et al. (1994) identify a five-phase model of analogy composed of *representation, retrieval, mapping, validation* and *induction*. We note a particular similarity between the last three phases of Wallas' model and the central three phases of Keane's model. That is, these phases involve finding inspiration, examining the implications of that inspiration and assessing its outcome. This paper concerns the use of a computational model of analogy, consisting of the three phases of *retrieval, mapping* and *validation*[2].

As we shall see, our three-phase model is capable of finding novel analogies and of generating novel inferences. In this paper our focus is on assessing the creative potential of this multi-phase model and to do this we provide it with a knowledge-base containing a variety of domain descriptions and examine the analogies that are generated.

To test the model's creative potential we decided to generate the maximum number of analogies that can be derived from a given set of domain descriptions. Of course in a more realistic scenario, one specific domain would probably be selected as a target problem, but there were two reasons for not doing so. First there was no reason to select one domain over all others as the target problem, particularly in the absence of other facets of intelligence or domain-specific expertise. Secondly, we are attempting to explore the creative potential of a three-phase model of analogy, we are not attempting to mimic the way one specific creative analogy was discovered.

The results described in Sections 5 and 6 of this paper were generated under the following scenario. The retrieval phase selects each domain from memory in turn, treating it as a target problem. Then for each of these targets, the model retrieves every other domain in turn and treats each as a candidate source. For each resulting analogy, the inter-domain mapping is generated and the resulting inferences were generated, evaluated and recorded. So for a memory containing *n* domains, the number of analogies generated is proportional to n^2.

3.1 The Three-Phase Analogy Model

While the focus in this paper is on assessing computer generated analogies, we now provide some details on the computational model itself. In principle however, any mapping and inference models could be used. First, retrieval was a simple exhaustive process that selected each domain and passed it as a candidate source to the (current) target.

The mapping model took each source and target in turn, identifying the inter-domain mapping for that analogical comparison. Like many mapping models (Keane and Brayshaw, 1988; Keane et al., 1994; Forbus et al., 1994; O'Donoghue et al., 2006), our mapping model followed the incremental mapping approach. Before being processed, the model identified the "level" of each relation. Relations taking objects as arguments were defined as level 1, a relation taking two level n relations as arguments was defined as level n+1.

The model then identified the "root" predicates in both the source and target domains (predicates forming the root of tree structures). A "root mapping" was identified between a root in the target and a predicate at the same level in the source - with a preference that this be a root predicate. These root mappings were then elaborated to include all compatible inter-domain mappings. Our mapping model employed Gentner (1983)'s predicate identicality constraint as a preference rather than a hard restriction.

Generating inferences followed the CWSG (Holyoak et al., 1994) pattern completion algorithm. These inferences were then validated using "functionally relevant attributes" (Keane, 1985; Eskridge, 1994) that were associated with each first-order relational predicate. A simple taxonomy supported this synta-semantic process. This validation process contrasts with the more detailed but domain-specific "verification" process used by Falkenhainer (1987).

Much of this paper is devoted to analyzing the computer generated analogies that were produced by our model. Our analyses focus on the issue as to whether our model does actually generate inferences, which display the hallmarks of creativity, such as *novelty* and *quality* (Colton and Steel, 1999; Ritchie, 2001). Rather than relying on intuition, we wish to statistically assess our model by examining the artefacts it produces.

3.2 Two Collections of Test Data

In order to test our model, we need domains that the model may process. Two distinct collections of domains were used to conduct two separate computational experiments. The first collection was developed by Veale (1995) and held 14 domains, each containing from 10 to over 100 predicates. The Professions domains contained descriptions of various professions (butcher, general, politician *etc*) and though it was designed to compare models of metaphoric mapping, this use has been extended in this paper. Each domain used between 6 and 15 distinct relational predicates (ignoring duplicates). So this collection consisted of large domains described using very general relational predicates - `(depend person personal-health)` and `(depend 18-th-century-general army)`.

The second collection was developed specifically for this project and contains 81 smaller domains, ranging in size from 1 to 15 predicates. These "Sundry domains" were described by much more specific relations - `(capture army fortress)` and `(bounce golf-ball golf-green)`. This collection also contained many domains found in the analogy literature, including Rutherford's *solar-system:atom* analogy and (Duncker, 1945) *tumour:fortress* analogy. These smaller domains used an average of M=3.48 distinct relational

[2] Validation differs from verification in that it is a more broadly applicable, but less intricate means of assessing the quality of analogies and their inferences.

predicates per domain.

3.3 Initial Testing

Initial testing of our model on a few domain descriptions revealed a number of findings. First, many analogical comparisons yielded no inferences. This occurred when no appropriate inter-domain mapping could be identified and when the source domain contained no additional material to be transferred to the target.

A second finding from our initial tests revealed that almost every inference generated was a novel inference. That is, almost none of the generated inferences were identical to a predicate already contained in the knowledge base. The majority of inferences (over 99%) were formed from a novel combination of a relational predicate plus its two arguments. While we have no measure of the degree of novelty of these predicates, such a low ratio of duplicates strongly indicates that the generated artefacts should be considered novel. To remove any nonsense inferences that might be generated *eg* `(sleep idea furiously)`, the validation model classified each inference as either valid or invalid. However, we must now focus on the task of assessing the goodness of our model. Was it successful in generating creative artefacts? Did the validation model successfully remove nonsense inferences? Was validation even necessary?

4 Assessing Novel Artefacts

Computational modelling must specify the processes and representations that underlie creativity, it must also generate creative artefacts. These artefacts must thus display the qualities associated with creativity: *quality, novelty* (Ritchie, 2001), *plausibility, surprisingness, applicability, utility etc.* (Colton and Steel, 1999). The main complication in assessing these qualities arises from the fact that these artefacts are also *novel* and this novelty has some surprising implications for how the other qualities may be assessed.

Firstly, we cannot use a direct comparison between the novel artefacts and some known set of artefacts (eg the inspiring set). Thus, assessment must be conducted in other terms. Gomes et al. (2003) assess the quality of novel artefacts in terms of the quantity of identified errors in the generated artefact. Veale (2004) compares the quality of generated artefacts to an independent resources (from WordNet). Falkenhainer (1987) "verifies" analogy-based physical models in terms of how well the new model matches (or can be adapted to) other known facts and rules. In this paper we present a more general approach to the analysis of creative artefacts. Like much of cognitive science we use human evaluators to assess the goodness of the artefacts produced.

4.1 Statistical Analysis

A common methodology in cognitive science is to examine people's performance at some task. Using this evidence and other information, an hypothesis (often instantiated as a computational model) is created of their performance at that task. The goodness of the hypothesis and the model is then assessed, often using parametric statistical techniques. Among the parametric statistics used are the Pearson product-moment correlation and ANOVA (analysis of variance) tests.

However, a number of differences mean that these statistical techniques can not be used to assess computational creativity. Firstly, we are not trying to compare the performance of a set of people to the model's performance at the same task. So, assumptions about the frequency distributions that underlie many of these statistical techniques do not hold. Secondly, cognitive science assesses how well a model accounts for observed phenomena. It does not normally attempt to identify specific qualities in computer generated items.

4.1.1 Non-Parametric Statistics

To assess our model we use non-parametric statistics. Non-parametric (or distribution free) statistics make no assumptions about the frequency distribution of the variables being assessed. Thus the model's structure is not specified beforehand but is derived from the data itself. While non-parametric tests have less power than parametric tests, they are generally more robust.

While it was intended to use (human) raters to asses the goodness of the generated artefacts *post hoc*, some additional constraints were also imposed by what can be expected of raters. Newly generated items were to be evaluated independently of the domain descriptions, because presenting raters with collection of up to 100 predicates was not thought likely to be successful. Our raters did not evaluate the analogical comparisons themselves, again as rating large pairings of predicates was considered too difficult. We evaluated the analogies indirectly, based on the inferences they mandated. Again inferences were evaluated in isolation and not as collections of predicates, partly because most inferences occurred as isolated predicates. Furthermore, assessing collections of predicates would require knowing the prior context – again involving reading larg(er) collections of predicates.

In this paper we make use of two different non-parametric tests; McNemar's test and the Mann-Whitney test. Within the context of this paper, the central difference between them is that the first test compares two binary classifications, while the second test compares a binary and an ordinal classification.

4.2 McNemar's Test

In this instance we use a McNemar's test to test for the presence of an hypotheses, that our model generated artefacts displaying some of the attributes of creativity (see Hinkle et al. (1994) for details on the McNemar's test). As stated earlier, virtually all inferences were already known to be novel. So, McNemar's test was used to assess the *quality* (Ritchie, 2001) of inferences. More specifically, we assess the validity of the analogically generated inferences. (In this paper evaluation is independent of the driving analogical comparison).

More specifically, this test will allow us to test the null hypothesis, that there will be an equal number valid inferences rated-bad and invalid inferences rated-good.

We start by recording the classifications assigned to each inference by our computational model. These inferences were then given to human raters for separate assessment, so the raters were unaware of the computers classification of these items. The assigned classes are then compared to the human ratings of these materials (see Table 1). What we would like is total agreement between the assigned classifications and the human ratings.

Table 1: Confusion Matrix of Results

Assigned Class	**Human Rating +**	**Human Rating -**	**Total**
Valid +	a	b	a + b
Invalid -	c	d	c + d

In assessing these data, McNemar's test focuses on the disagreement between the categorisation and the human rating (Hinkle et al., 1994).

$$\chi^2 = \frac{(b-c)^2}{(b+c)} \tag{1}$$

McNemar's test will help us decide if our model produces valid artefacts that people think of as valid. That is, people agree that what the model categorises as a valid inference is indeed a valid inference. People also agree that the invalid inferences are invalid. Thus, the quality of generated artefacts is assessed in terms of their validity (novelty being assessed independently).

4.3 Mann-Whitney-Wilcoxon Test

To further analyse our results, a Mann-Whitney(Wilcoxon) test was also performed on the data. The Mann-Whitney-Wilcoxon test improves upon the McNemar's test by taking into account the ordinal scale used by the human raters to rate the novel artefacts. (Thus, the McNeemar's test is included in this paper for illustrative purposes). The Mann-Whitney Test is one of the most powerful of the nonparametric tests.

Mann-Whitney tests assesses if two samples come from the same distribution. The null hypothesis is that the two samples are from the same population and that their probability distributions are the same.

The two categories (valid and invalid) are combined and sorted by their rating score. The combined data are ranked and rank-sum for each category is computed (R_1 and R_2). Tied results are given the average value for that ranking group. Equation 2 details how U_1 value is calculated (an equivalent equation exists for U_2 and R_2) with U being chosen as the smaller of U_1 and U_2.

$$U_1 = mn + \frac{m(m+1)}{2} - R_1 \tag{2}$$

where m and n are the numbers of items in the two categories. For large sample sizes (n>20) an approximation can be used. Additionally, because of the presence of a large number of tied rankings among our results, we made use of a further modification to the basic formula that includes a correction factor to account for the presence of these tied rankings. Further details on the Mann-Whitney test can be found in (Siegel and Castellan, 1988).

$$\frac{z = W_x + .5 - n(N+1)/2}{[mn/N(N-1)][(N^3-N)/12 - \sum_{j=1}^{g}(t_j^3 - t_j)/12]} \tag{3}$$

where N=m+n, t_j is the number of tied ranks in the jth grouping, W_x is the sum of the ranks for the first category and g is the number of groupings of different tied ranks.

5 Analysis of Results

In this section we describe the results of a number of tests that were conducted on our model. A large number of analogies were generated and then assessed by examining their inferences. The quality of the resulting inferences is examined using the tests mentioned above. Additionally, some factors relating to the representation of information arise from these results, so some facets of the domain descriptions are also examined.

5.1 Experimental Set-Up

A memory was created containing all domains from two knowledge bases (described below). Each of these domains were taken in turn to serve as the target problem. Every domain was taken in turn to act as a candidate source for that target and the inter-domain mapping and inferences were generated (Holyoak et al., 1994). These inferences were then passed to a validation process, which categorised all inferences as either *valid* or *invalid*.

5.1.1 Participants and Design

Two raters were used and both raters were familiar with predicate calculus representation. All data were presented in a random order.

5.1.2 Procedure

Unrated inferences were given to human subjects who were asked to give each predicate a rating between 1 and 7. A rating of 1 represented a predicate that could not be considered credible under any circumstance, while a rating of 7 represented a predicate that could certainly be considered credible in some circumstance. A rating of 4 represented a predicate that was not obviously either credible or not credible in any circumstance.

The materials used for Experiment 1 was the inferences generated on the Professions domains. The materials used for Experiment 2 were the inferences generated on the Sundry domains. The same experimental set-up was used to produce all results.

5.2 McNeemar's Analysis

In this section we present the results of a McNeemar's analysis of the experimental data. We first present the results for the Professions domains. Next we present the results from the Sundry domains and then we compare the two results. The 7 point rating was then mapped onto a binary scale of Rated-valid or Rated-invalid, for use in the McNemar's test.

5.2.1 Experiment 1

The 14 domains from the first collection generated 196 analogies, representing each domain mapped with all other domains - including itself. The model generated a total of 175 inferences and classified 151 (86.2%) as *valid*, and 24 (13.7%) as *invalid*. Of the 175 generated inferences, 40 (approximately 1/4) were randomly selected for rating.

The average rating awarded to predicates that the model categorised as *valid* was M=2.77 (SD=1.98), while the average rating awarded to the *invalid* predicates was M=1.58 (SD=1.06). So as expected, the valid predicates were generally rated better than the invalid predicates.

Of the 20 *valid* predicates, 6 (30%) were rated as valid or potentially valid (rated >=4) by the raters, so 14 (70%) of the *valid* category were actually deemed invalid by the human raters. Of the 20 *invalid* predicates 19 (95%) were rated as invalid and 1 (5%) was rated as valid. Thus, the model appears to be better at identifying invalid predicates than it is at recognising valid predicates. This may be explained by the fact that predicates are only categorised as invalid when some specific violation of the functionally relevant attributes is identified. Otherwise, predicates are assumed to be valid.

Table 2: Assessing Generated Analogies - Collection 1

Assigned Class	Rated Valid	Rated Invalid	Total
Valid	3(20%)	12(80%)	15(100%)
Invalid	1(4.2%)	18(94.7%)	19(100%)

The first assessment of our computer generated items is summarised by a McNemar's test. The McNemar's test compared the classifications of the computer generated items to categorisations awarded by human raters to the same items. In this case the null hypotheses states there will be an equal number of inferences in the Invalid-RatedGood and the Valid-RatedBad conditions. The results were: #Invalid-RatedGood = 1, #Valid-RatedBad = 14, K^2=11.26 and taking α = 0.05 the null hypothesis can be rejected. p $<=$ 0.001 showing strong agreement between the two ratings, indicating that the model correctly interpreted its own output. Thus, the model was successful in generating quality artefacts that were judged to be valid by human raters.

5.2.2 Experiment 2

The second collection of 81 domains generated a total of 6561 analogies, yielding 3793 inferred predicates. Of these predicates, 2158 (56.9%) were classified as *valid* and 1635 (43.1%) inferences were categorised as *invalid* predicates.

216 valid predicates and 50 invalid predicates were randomly selected for human rating (these quantities being related to the technique which ensured a random selection was made). Of the 216 *valid* predicates, 103 (47.5%) were rated as valid or potentially valid by the raters, so 94 (43.5%) of the *valid* category were actually deemed invalid by the human raters. Of the 50 *invalid* predicates 45 (90%) were rated as invalid and 5 (10%) were rated as valid.

The average rating awarded to the *valid* predicates was M=3.47 (SD=2.35), for the *invalid* predicates was M=1.59 (SD=1.42). Thus as expected, the invalid predicates received significantly lower ratings than the valid predicates. As with Experiment 1, the invalid category is recognised with greater accuracy than the valid category.

Table 3: Assessing Generated Analogies - Collection 2

Assigned Class	Rated Valid	Rated Invalid	Total
Valid	94(43.5%)	122(56.5%)	216(100%)
Invalid	5(14%)	45(86%)	50(100%)

A McNemar's test was also performed to compare the model's classifications to the categorisations awarded by the raters. The results were: #Invalid-RatedGood = 5, #Valid-RatedBad = 122, K^2= 107.78, α = 0.05 so again the null hypothesis can be rejected. p $<=$ 0.0001 showing a very strong agreement between the ratings and the assigned category.

5.2.3 Discussion on Experiments 1 and 2

McNemar's test allows us to reliably reject the null hypothesis. However, a comparison between the two experiments provides greater insight.

The validation model is being very cautious about categorising relations as invalid, only doing so when there is reasonable evidence. If there is doubt about a relation's validity, it is passed as potentially valid. Thus, inferences assigned to the valid class consist of true valid inferences as well as invalid inferences on which there was insufficient information.

The average rating for the *valid* inferences in the first collection (M=2.77, SD=1.98) was significantly lower than the second (M=3.47, SD=2.35). Thus, inferences were validated less successfully on the first collection than on the second collection. However, the proportion of Valid-RatedGood = 20% in the first collection was significantly lower than on the second collection Valid-RatedGood = 43%. This can be attributed to the fact that the first collection used more general relational predicates, which are more difficult to falsify. Secondly, the second collection made greater use of relational predicates defined by functionally relevant attributes that supported the classification process.

In conclusion, it appears that the validation process is primarily responsible for the quality of inferences in the valid and invalid categories. Domains that are described using more specific relations (from lower-down a taxonomy) allow the validation process to operate more accurately.

5.3 Mann-Whitney(Wilcoxon) Analysis

A Mann-Whitney analysis waas conducted on our results. As stated earlier, the main results below counter for the presence of large number of tied results. The presence of

tied rankings was a greater factor in the analogies generated from the second collection than on the first collection of domains.

We first present the results for the Professions domains and then results from the Sundry domains before comparing the two results.

The materials and method used in this experiment were the same as in the previous analysis.

The null hypothesis tested in this section is that the two samples are from the same population and that their probability distributions are the same.

5.3.1 Experiment 3

With the formula given above, the results for our Mann-Whitney test are : R_1 323.5, R_2 498, U = 112, z=2.3 (p_1 <0.0107, p_2 <0.0214).

However, when we use the Mann-Whitney test that is adjusted for the presence of many tied results. z=2.49, (p_1 <0.0064, p_2 <0.02).

This result allow us reject the null hypothesis, that the valid and invalid categories are drawn from the same population. We can thus adopt the alternate hypothesis that the mean of the valid category is greater than the invalid category. Thus our analogy model is indeed generating quality analogical inferences.

5.3.2 Experiment 4

With the formula given above, the results for our Mann-Whitney test are : R_1 30057.5, R_2 5414, U = 4147.5, z=2.55 (p_1 <0.005, p_2 <0.0108).

However, if we include the correction factor to account for the presence of tide results in our ranking, then z = 5.92 (p_1 <0.0001, p_2 <0.0001).

Thus we reject the null hypothesis in favour of the alternate hypothesis, which is that the median of the valid inferences is greater than the median of if invalid inferences. Alternatively we may state that the valid inferences have stochastically greater ratings than the invalid inferences.

Again this was the hoped for result and shows that our analogy model does generate quality inferences.

5.3.3 Discussion on Experiments 3 and 4

As expected, these results indicate that the null hypothesis can be rejected. Again, the results from the Assorted collection given graeter confidence for this conclusion than do the Professions results.

5.4 Known Creative Analogies

For clarity, we shall report separately on the accuracy of our model to re-generate known creative analogies. As discussed in Section 2.1 we do not consider the following results to be examples of true creativity - because the domain descriptions do not accurately reflect the problems that were creatively solved in each instance. However, they do provide some positive evidence for the creative potential of the analogy model.

Because the creative analogies were known *a priori*, we do not need to use McNemar's test. The model generated and validated the correct inferences for 7 of 10 (70%) (known) creative analogies and thus was quite successful in (re)generating these analogies.

While no new creative analogies were discovered on these knowledge-bases, we believe that creative analogies could be discovered by our model. These results show us that creative analogies occur exceptionally rarely. A challenge for the future is to acquire more domain descriptions to see if any creative analogies are generated. A related challenge is to improve the evaluation process in order to focus on the more promising and creative analogical comparisons.

6 Conclusion

The traditional approach to computational creativity attempts to generate new items belonging to an "inspiring set". But this inspiring set also plays a role in evaluating the creativity model.

We describe an alternative *process-centric* approach to computational creativity that does not utilise an inspring set. Thus, evaluating these models must rely on alternative methods. This paper describes two nonparametric statistics techniques, namely McNeemar's test and the Mann-Whitney test. These tests evaluate artefacts that have been rated on binary and ordinal scales respectively.

These statistical techniques were used to evaluate a model of analogical reasoning that has been adapted to operate as an analogy generating machine. This model encompassed the three core phases in the analogy process, namely: *retrieval*, *mapping* and *validation*. The model was used on a knowledge base containing two distinct collections of domains, to assess its performance and see if any novel or creative analogies might be generated. Every domain was used as a target in conjunction with each other candidate source domain. This generated the maximum number of analogical inferences allowing us to test the creative potential of our model.

The resulting inferences were evaluated by the model itself, selecting inferences of greater quality. These inferences were recorded and given to human raters who assessed the accuracy of the analogy production system. A McNemar's test was used to compare and assess the automatically assigned classification against human ratings of the same artefacts. This illustrated that the model was successful in generating quality inferences.

Known examples of creative analogies were identified as expected (such as Rutherford's *solar-system:atom* analogy). However, such examples are not considered as *truly* creative as they have been described in such a way as to maximise the similarity between the two domains making their re-discovery almost inevitable. No *truly* creative analogies were identified among the subset of inferences rated by the raters.

Interesting differences between the two collections produced differences in the generated results. The collection using more general (or abstract) relational-predicates made generating the mapping easier, but made validation less accurate. In contrast, the collection using more specific relational-predicates made identifying the inter-domain mapping more difficult, but allowed more accurate validation of inferences.

As far as we know, this is the first work towards automatically generating analogies. While the analogies reported in this paper were not found to be creative, we believe a larger knowledege-base will provide more fruitful results. More accurate and complete models of each phase of analogy may help further improve the quality of results produced by the model. Modifying the model's parameters may even produce a model with a greater creative capacity than human analogisers.

Acknowledgements

My sincere thanks to the anonymous reviewers of IJWCC 2007 (one in particular) for helpful comments and suggestions on this paper.

References

Boden, M. A. (1992). *The Creative Mind*. Abacus.

Brown, T. L. (2003). *Making Truth: Metaphor in Science*. University of Illinois Press, New York, USA.

Colton, S. and Steel, G. (1999). Artificial intelligence and scientific creativity. *Artificial Intelligence and the Study of Behaviour Quarterly*, 102.

Curie, M. (1923). *Pierre Curie*. Macmillan.

Dunbar, K. (2001). *The Analogical Mind*, chapter The Analogical Paradox: Why Analogy is so Easy in Naturalistic Settings, Yet so Difficult in the Psychological Laboratory.

Dunbar, K. and Blanchette, I. (2001). The in vivo/in vitro approach to cognition: The case of analogy. *Trends in Cognitive Sciences*, 5(8):334–339.

Duncker, K. (1945). On problem solving. *Psychological Monographs*, 5:whole no. 270.

Eskridge, T. C. (1994). *Analogy, Metaphor and Reminding*, chapter A Hybrid Model of Continuous Analogical Reasoning. Ablex, Norwood, NJ.

Eysenck, M. W. and Keane, M. T. (1995). *Cognitive Psychology*. Taylor Francis, Erlbaum, UK.

Falkenhainer, B. (1987). An examination of the third stage of the analogy process: Verification-based analogical learning. In *Proc. IJCAI*, pages 260–263.

Falkenhainer, B., Forbus, K., and Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence*, 41:1–63.

Forbus, K., Ferguson, R., and Gentner, D. (1994). Incremental structure-mapping. In *Proc. 16th Cognitive Science Society*, pages 313–318.

Forbus, K., Gentner, D., and K., L. (1995). Mac/fac: A model of similarity-based retrieval. *Cognitive Science*, 19:141–205.

French, R. M. (2002). The comptational modeling of analogy-making. *Trends in Cognitive Sciences*, 6(5):200–205.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7:155–170.

Gomes, P., Seco, N., Pereira, F. C., Paiva, P., Carreiro, P., Ferreira, J. L., and Bento, C. (2003). The importance of retrieval in creative design analogies. In *Proc. IJCAI 3rd Workshop on Creative Systems*, Acapulco, Mexico.

Hinkle, D. E., Wierrsma, W., and Jurs, S. G. (1994). *Applied Statistics for the Behavioral Sciences*. Houghton Mifflin, Boston, USA.

Hoffman, R. (1995). Monster analogies. *AI-Magazine*, 3:11–35.

Holyoak, K., Novick, L., and E., M. (1994). *Analogy, Metaphor and Reminding*, chapter Component Processes in Analogical Transfer, pages 113–180. Ablex Norwoord, N.J., USA.

Keane, M., Ledgeway, T., and Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, 18:387–438.

Keane, M. T. (1985). On drawing analogies when solving problem:. *British Journal of Psychology*, 76:449–458.

Keane, M. T. and Brayshaw, M. (1988). *Third European Working Session on Machine Learning*, chapter Indirect Analogical Mapping. Pitman, London, UK.

Koestler, A. (1964). *The Art of Creation*, volume 1. Picador, London.

O'Donoghue, D. (1997). Towards a computational model of creative reasoning. In *Conference on Computational Models of Creative Cognition (CMOCC)*, Dublin City University, Ireland.

O'Donoghue, D. and Crean, B. (2002). Retrieving creative analogies. In *ECAI - Workshop on Creative Systems*, pages 56–66, Lyon, France.

O'Donoghue, D. P., Bohan, A., and Keane, M. (2006). Seeing things : Inventive reasoning with geometric analogies and topographic maps. *New Generation Computing*, 24:267–288.

Plate, T. (1998). Structured operations with distributed vector representations. In *Advances in Analogy Research*, Sofia, Bulgaria,.

Ritchie, G. (2001). Assessing creativity. In *Proc. AISB Symposium on AI and Creativity*, pages 3–11, York, England.

Siegel, S. and Castellan, J. N. (1988). *Nonparametric Statistics*. McGraw-Hill.

Thagard, P., Holyoak, K. J., Nelson, G., and Gochfeld, D. (1990). Analogue retrieval by constraint satisfaction. *Artificial Intelligence*, 46:259–310.

Veale, T. (1995). *Metaphor, Memory and Meaning*. Ph.d. diss., Trinity College, Dublin.

Veale, T. (2004). Pathways to creativity in lexical ontology. In *Proc. 2nd Global WordNet Conference*.

Wallas, G. (1926). *The Art of Thought*, volume 1. Cape, London.