Baseline Methods for Automated Fictional Ideation

Maria Teresa Llano, Rose Hepworth, Simon Colton, Jeremy Gow and John Charnley
Computational Creativity Group, Department of Computing, Goldsmiths, University of London

Nada Lavrač, Martin Žnidaršič and Matic Perovšek
Department of Knowledge Technologies, Jožef Stefan Institute

Mark Granroth-Wilding and Stephen Clark
Computer Laboratory, University of Cambridge

Abstract

The invention of fictional ideas (ideation) is often a central process in the creative production of artefacts such as poems, music and paintings, but has barely been studied in the Computational Creativity community. We present here three baseline approaches for automated fictional ideation, using methods which invert and alter facts from the ConceptNet and ReVerb databases, and perform bisociative discovery. For each method, we present a curation analysis, by calculating the proportion of ideas which pass a typicality evaluation. We further evaluate one ideation approach through a crowdsourcing experiment in which participants were asked to rank ideas. The results from this study, and the baseline methods and methodologies presented here, constitute a firm basis on which to build more sophisticated models for automated ideation with evaluative capacity.

Introduction

Ideation is a portmanteau word used to describe the process of generating a novel idea of value. Fictional ideation therefore describes the production of ideas which are not meant to represent or describe a current truth about the world, but rather something that is in part, or entirely, imaginary. As such, their purposes include unearthing new truths and serving as the basis for cultural creations like stories, advertisements, poems, paintings, games and other artefacts. Automated techniques for the derivation of new concepts have been important in Artificial Intelligence approaches, most notably machine learning. However, the projects employing such techniques have almost exclusively been applied to finding concepts which somehow characterise *reality*, rather than some fictional universe. While some concepts may be purported as factual, i.e., supported by sufficient evidence, others may only be hypothesised to be true. In either case, however, the point of the exercise is to learn more about the real world through analysis of real data, rather than to invent fictions for cultural consumption.

A major sub-field of Computational Creativity research involves designing software that exhibits behaviours perceived as creative by unbiased observers (Colton and Wiggins 2012). However, in the majority of the generative systems developed so far within Computational Creativity research, there is no idea generation undertaken explicitly. An exception to this was (Pereira 2007), who implemented a system based on the psychological theory of Conceptual Blending put forward by Fauconnier and Turner (2008). By blending two theories about different subject material, novel concepts which exist in neither domain emerge from the approach. Using blending to reason about such fictional ideas was harnessed for various creative purposes, including natural language generation (Pereira and Gervás 2003), sound design (Martins et al. 2004), and the invention of character models for video games (Pereira and Cardoso 2003). Similarly, the ISAAC system (Moorman and Ram 1996) implements a theory for creative understanding based on the use of an ontology to represent the dimensions of concepts. By altering the dimensions of existing concepts within the ontology, for instance considering a temporal object as a physical one, the system is able to create novel concepts.

In addition, in some projects, especially ones with application to natural language generation such as neologism production (Veale 2006), which are communicative in nature, it is entirely possible to extract ideas from the artefacts produced. However, it is fair to say that such software is not performing ideation to produce artefacts, but is rather producing artefacts that can be interpreted by the reader via new ideas. The work in (Goel 2013) shows the use of creative analogies in which problems of environmental sustainability are addressed by creating designs inspired by the way things work in nature. For instance, birds' beaks inspired the design of trains with noise reduction. Although ideation here is being used for inspiration and not to create literal representations, this work shows the potential of using creative analogies for fictional ideation.

As part of the WHIM project[1] (an acronym for the *What-if Machine*), we are undertaking the first large-scale study of how software can invent, evaluate and express fictional ideas. In the next section, we present three straightforward approaches to fictional ideation which manipulate material from internet sources. These will act as our baseline against which more sophisticated ideation methods will be tested as the project progresses. In order to draw that baseline, we conducted a *curation analysis* of the ideas produced by each method, whereby we calculated the proportion of ideas which were typical in the sense of being both understandable and largely fictional, with details given below. We also

[1] www.whim-project.eu

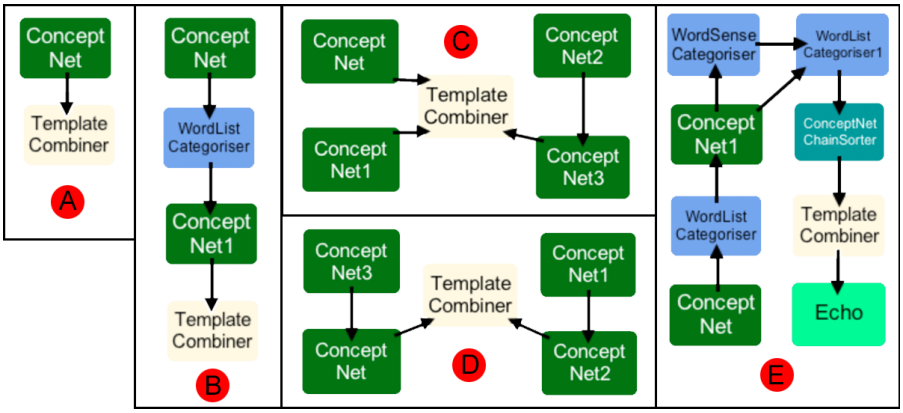

Figure 1: Ideation flowcharts using ConceptNet.

present here a baseline methodology for estimating the true value of the ideas produced by our systems. To do this, we conducted a crowd-sourcing exercise involving 135 participants, where people were exposed to ideas in a controlled way, with the aim of evaluating components of ideas that could be used to predict overall value.

A good fictional idea distorts the world view around it in useful ways, and these distortions can be exploited to spark new ideas, to interrogate consequences and to tell stories. A central hypothesis of the WHIM project is that the narrative potential of an idea can be estimated automatically, and used as a reliable estimate of the idea's worth. Hence the crowd-sourcing study had narrative potential as a focal point, and we tested an automated approach which estimates whether an idea has much narrative potential, or little. As discussed below, we found that, in general, people ranked those ideas that were assessed as having much potential higher than those assessed as having little. We present further statistical analysis of the results, which enables us to conclude by describing future directions for the WHIM project.

Baseline Ideation Methods

We investigate here three methods which use data mined from the internet for generating What-if style fictional ideas. In the next section, we analyse the results from each method.

Fictional Ideation using ConceptNet

ConceptNet[2] is a semantic network of common sense knowledge produced by sophisticated web mining techniques at the MIT media lab (Liu and Singh 2004). Mined knowledge is represented as facts, which comprise relations between concepts in a network-like structure, e.g., [*camel, IsA, animal, 7.0*], [*animal, CapableOf, hear_sound, 2.0*]. Currently, ConceptNet has 49 relations, including UsedFor, IsA, AtLocation, Desires, etc., and each fact is given a score, from 0.5 upwards, which estimates the likelihood of the relation being true, based on the amount of evidence mined. We have studied fictional ideation by *inverting* the world view modelled by ConceptNet, i.e., facts are transformed by negating their relations. For example, this can be done by introducing an action which was not previously possible, e.g., 'people can't fly' becomes *What if people could fly?* or stopping an action or desire which was previously common, e.g., 'people need to eat' becomes *What if people no longer needed to eat?*, etc. We investigated various inversion methods such as these, carried out using the FloWr flowcharting system described in (Charnley, Colton, and Llano 2014).

Working in a story-generation context, we took inspiration from the opening line of Franz Kafka's 1915 novella *The Metamorphosis*:

> "One morning, as Gregor Samsa was waking up from anxious dreams, he discovered that in his bed he had been changed into a monstrous verminous bug".

In figure 1, we present five flowcharts we used to generate ideas by inverting and combining ConceptNet facts about people, animals, vegetables and materials.

Flowchart A finds instances of animals by searching ConceptNet for facts [X, IsA, animal]. These are then rendered in the *TemplateCombiner* process as questions of the form: "What if there was a person who was half man and half X?" Flowchart B employs ConceptNet similarly, then uses a *WordListCategoriser* process to remove outliers such as [my_husband,IsA,animal]. Then, for a given animal, A, facts of the form [A,CapableOf,B] are identified and rendered as: "What if there was a person who was half man and half X, who could Y?" Switching the CapableOf relation to NotCapableOf enabled us to produce ideas suggesting a person who became an animal, but retained some human qualities. We augmented this by using the LocatedNear relation (not shown in figure 1) to add a geographical context to the situation, producing ideas such as "What if a woman awoke in the sky to find she had transformed into a bird, but she could still speak?" We found that these ideas had much resonance with the premise in *The Metamorphosis*.

Taking our lead next from the surrealistic artworks of Dali, Magritte and colleagues, in flowchart C, we looked at bizarre visual juxtapositions. ConceptNet is used here to find an occupation, a vegetable and a location related to some animal, and the flowchart produces ideas such as: "What if there was a banker underwater with a potato for a face?" Similarly, in flowchart D, we produced ideas for paintings by finding materials, M, using facts of the form

[2] conceptnet5.media.mit.edu

[X,IsA,thing] and [X,MadeOf,M], then finding organisms, O, with pairings of [X,IsA,live_thing] and [O,IsA,X] facts. This led to ideas such as painting a dolphin made of gold, a reptile made of wood, and a flower made out of cotton. In the baseline evaluation section below, we describe the raw yield of flowcharts A to D, and the proportion of the results which were both understandable and mostly fictional.

As mentioned above, we are particularly interested in estimating the *narrative potential* of an idea, by which we mean the likelihood that the idea could be used in multiple, interesting and engaging plots for stories. As a baseline method for estimating such potential, we investigated a technique consisting of building inference chains of ConceptNet facts whose starting point is the fact that is inverted in the idea. To illustrate the approach, from the seed idea *"What if there was a little bug who couldn't fly?"*, the following chain of relations can be obtained through ConceptNet:

[bug,CapableOf,fly] → [fly,HasA,wing] → [wing,IsA,arm] → [arm,PartOf,person] → [person,Desires,muscle] → [muscle,UsedFor,move_and_jump]

Here, one can imagine a bug who can't fly, but instead uses his muscle-bound human like arms for locomotion.

Our hypothesis is that, while each chain might be rather poor and difficult to interpret as a narrative, the volume and average length of such chains can indicate the potential of the idea. We implemented a *ConceptNetChainSorter* process to take a given idea and develop chains up to a specified length with no loops or repetitions. Flowchart E uses this process to order the facts from ConceptNet in terms of the sum of the lengths of the chains produced. Hence facts with many chains are ranked higher than chains with fewer, and longer rather than shorter chains will also push a fact up the rankings. Often there are no chains for a fact, and if there are, the number depends on the nature of the objects being related, and the relation. Looking at facts [X,R,Y], where [X,IsA,animal] is a ConceptNet fact, for each R, we found these percentages of facts had non-trivial chains:

CapableOf	Desires	HasA	HasProperty	IsA	LocatedNear
20	50	63	28	48	100

Fictional Ideation using ReVerb

The Washington ReVerb project (Fader, Soderland, and Etzioni 2011) extracts binary relationships between entities from text, like the ConceptNet relations described above. Output produced by running the system over a large corpus of web texts (ClueWeb09, ∼1 billion web pages) is publicly available and we use it here to generate fictional ideas. Lin, Mausam, and Etzioni (2012) have linked the first argument (LHS concept) of a subset of ReVerb extractions with identifiers of an entity in Freebase (Bollacker et al. 2008). This provides a means of unifying the various names by which a particular entity might be referred to (*cow, cattle*, etc.) and disambiguating entities that have the same name. In the ideation method described here, we use this dataset, and the input to the process is a Freebase ID.

The relations vary in generality, as well as reliability. For example, some relations express a particular one-off event during which the entities interacted (*Tony Blair converted to Catholicism*), while others express general properties of the entities (*cows eat grass*). Both types of relations may be of interest to building world views for ideation, and we do not attempt to distinguish them currently. Using facts from ReVerb, we can generate fictional ideas by substituting one of the arguments for an alternative entity. For example, the extractions relating to *cattle* include *[Cattle, were bred for, meat]*. Looking at other facts that use the same relation (*be bred for*), with different LHS entities, we find things that are bred for *speed*, suggesting a possible fictional fact: *[Cattle, were bred for, speed]*.

The following are desirable properties of such alterations:

1. They should be fictional (e.g., *[Cattle, were bred for, meat]* $\not\Rightarrow$ *[Cattle, were bred for, milk]*).
2. They should make sense (e.g., *[Cattle, were bred for, meat]* $\not\Rightarrow$ *[Cattle, were bred for, rule of thumb]*).
3. They should have a substantial effect on the narratives that could be generated (e.g., *[Cattle, were bred for, meat]* $\not\Rightarrow$ *[Cattle, were bred for, hamburgers]*).

Establishing whether this last desideratum holds is a hard task which we leave for now to future work.

Given an extraction $[X, r, Y]$, we wish to generate a fictional $[X, r, Y']$. The following requirements might serve to approximate the first two desiderata above:

- $[X, r, Y'']$ is common for some Y'', i.e., r is a common type of fact to say about X.
- $[X', r, Y']$ is common, i.e., Y' is commonly seen as the second argument of r (with different first arguments).
- $[X, r, Y']$ is rarely or never seen, i.e., this is likely not a fact we are already aware of. As we cannot rely on the dataset to contain all relevant facts, we impose a strong version of this, that $[X, r, Y']$ is completely unattested.

As an example, the following alteration is well supported by these criteria: *[Michael Jackson, was still the king of, pop]* ⇒ *[Michael Jackson, was still the king of, Kong]*. The initial fact is chosen because Michael Jackson is frequently said to have been the king of things (popular music, music video, etc.) – the first requirement. Kong is chosen as an alternative second argument, because Kong ranks highly among things that people are described as being still king of[3] – the second requirement. Finally, we have never seen Michael Jackson described as being still the king of Kong.

The first two requirements given above can be expressed, and combined, as conditional probabilities. $P(r|X)$ represents the probability of the relation given the first argument (the input). This will be high for the relations most often seen with X as the first argument (the most common things to say about X). $P(Y'|r)$ will likewise be high for the most common second arguments of the relation in question, regardless of which X they have been seen with. To eliminate attested facts, we exclude any Y' seen at all in $[X, r, Y']$. For each of the top 100 facts about X found in the ReVerb extractions, all alterations Y' with a non-zero $P(Y'|r)$ are ranked according to $P(r|X) \times P(Y'|r)$.

[3] High-scorers in the game Donkey Kong are described as such.

Below are some examples of the alterations the system performs, with an analysis of the proportion of usable alterations given in the next section. The following are the top five alterations for entity *cattle*, showing the fact in its extracted form, then the system's alteration, which could be rendered as a What-if style idea:

1. Cattle evolved to eat grass ⇒ Cattle evolved to eat meat
2. Cattle occupy a unique role in human history ⇒ Cattle occupy a unique role in Israelite history
3. Cattle occupy a unique role in human history ⇒ Cattle occupy a unique role in modern distributed systems
4. Cattle occupy a unique role in human history ⇒ Cattle occupy a unique role in society
5. Cattle were bred for meat ⇒ Cattle were bred for speed

Similarly, the top five for *Scotland* are:

1. Scotland is steeped in history ⇒ Scotland is steeped in tradition
2. Scotland is a part of the United Kingdom ⇒ Scotland is a part of life
3. Scotland is in Britain ⇒ Scotland is in trouble
4. Scotland is in Britain ⇒ Scotland is in order
5. Scotland is in Britain ⇒ Scotland is in progress

In other tests, we produced ideas that express fictional histories, which is a mainstay of creative writing, for instance: "What if John F. Kennedy had been elected Pope?"

Fictional Ideation using Bisociative Discovery

Koestler (1964) stated that different types of invention all share a common pattern, to which he gave the term "bisociation". According to Koestler, bisociative thinking occurs when a problem, idea, event or situation is perceived simultaneously in two or more "matrices of thought" or domains. When two matrices of thought interact with each other, the result is either their fusion in a novel intellectual synthesis, or their confrontation in a new aesthetic experience.

The developers of the CrossBee system (Juršič et al. 2012) followed Koestler's ideas by exploring a specific form of bisociation: finding terms that appear in documents which represent bisociative links between concepts of different domains, with a term ranking method based on the voting of an ensemble of heuristics. We have extended this methodology with a banded matrices approach, described in (Perovšek et al. 2013), which is used in a new CrossBee heuristic for evaluating terms according to their bridging term (b-term) potential. The output from CrossBee is a ranked list of potential domain bridging terms. Inspecting the top-ranked b-terms should result in a higher probability of finding observations that lead to the discovery of new links between different domains. Here, the creative act is to find the links which cross two or more different domains, leading out of the original 'matrix of thought'.

In the simplified ideation scenario addressed here, we used CrossBee for b-term ranking on documents from two domains to discover bridging terms, with the aim of combining statements from two domains. The first domain consists of 154,959 What-if sentences retrieved from Twitter with query 'what if', assisted by the Gama System® PerceptionAnalytics platform.[4] The tweets were filtered through the following steps, reducing the number to 65,811:

- All non-ASCII characters were deleted.
- Repeated letters were truncated, so that any character repeating consecutively more than twice in a word was ignored after the second repetition. For example, the word *cooooool* would be truncated to *cool* (but also *looooooove* would be truncated to *loove*).
- All characters are transformed to lower case.
- Non-English tweets were removed.
- Vulgar words were removed by comparison with a list of such words.[5]
- From all items, only the sub-strings starting with the term 'what if' and ending with a period, question mark or exclamation mark were considered.
- Items shorter than 9 characters were removed.
- Exact duplicates were removed.

The second dataset is a collection of 86 moral statements from Aesop's fables, which was created by crawling the Aesop's fables online collection. Each What-if sentence and each moral statement was treated as a separate document, and all documents were further preprocessed using standard text mining techniques. We then applied our methodology to the data from the two domains to estimate the b-term potential of common terms. We used this indicator for ranking (a) single What-if sentences and (b) bisociatively linked What-if sentences and moral statements.

Inspection of the What-if sentences obtained from tweets revealed that a great number of them make very little sense in general or are related to very specific contexts. Aesop's morals, on the other hand, tend to be very general in nature. By composing sentences from these two domains using the terms with the best b-term potential indicator value, we hoped to produce a ranking mechanism that favours generally meaningful fictional ideas that might be useful for ranking individual What-if sentences. We used the mechanism to rank both single sentences and compound pairs, to test the hypothesis that using the b-term potential as an ranking coefficient can estimate which What-if sentences will be evaluated more favourably by people, both as individual sentences and in bisociatively combined sentence pairs.

The effectiveness of b-term potential used as a ranking tool of single What-if sentences was evaluated as follows: we randomly shuffled the 10 best ranked sentences and 10 random What-if sentences. The collection of 20 sentences was then independently assessed by 6 human evaluators who used scores from 1 (bad) to 5 (very good) in answering the question: "*How good (generally interesting) do you find the following idea?*" The top 10 b-term ranked What-ifs received an average score of 2.92, whereas the randomly chosen ones scored 2.80 on average. Application of an Unpaired

[4] demo.perceptionanalytics.net

[5] urbanoalvarez.es/blog/2008/04/04/bad-words-list/

T-test suggests that the difference among these two scores is not significant (p=0.6736). The best ranked What-if, according to the b-term potential was: "*What if a called myself the pope then charged into the vatican and demanded a duel to the death with an old man?*" This was also the sentence that achieved the best average score from the human evaluators.

The impact of b-term potential ranking on compound sentence pairs was evaluated similarly. To do this, we took the top 4 What-ifs and the top 4 moral statements that contained the strongest b-term. By combining them, we created a collection of 16 pairs of sentences. This collection was compared to two other collections: (i) a collection of 16 pairs of sentences (What-if + moral) that shared a b-term regardless of its strength, and (ii) a collection of 16 randomly paired What-if and moral sentences. Our hypothesis was that the top ranked collection will score higher on average than the one with randomly ranked b-terms and significantly better than the one which was randomly put together, ignoring b-terms. The pairs were randomly shuffled and independently assessed by 6 human evaluators answering the question: "*How good do you find the combination of the two sentences?*", scored again from 1 to 5.

Surprisingly, the top ranked collection was scored significantly (p=0.0076) lower than the randomly ranked one, with average score of 2.43, compared to 2.96. Also, in an independent comparison, it scored lower than the randomly paired sentences, having an average score of 2.70 compared to 2.78, although this was not significantly lower (p=0.6677). The compound sentence pair with the best b-term rank was: "*What if a called myself the pope then charged into the vatican and demanded a duel to the death with an old man? Every man should be content to mind his own business*". However, this sentence pair was ranked only 8th best among 32 manually evaluated compound pairs.

Given the encouraging result of the ranking mechanism for single What-if sentences, and the bad performance on its target compound data, the usefulness of the bisociative discovery methods for ideation and idea assessment cannot be confirmed. Hence, we plan further implementation and experimentation. In particular, we will enlarge the dataset of moral statements, to strengthen the bisociation approach.

Curation Analyses

Recall that we plan to use the above ideation methods as a baseline against which to compare more sophisticated approaches as the WHIM project progresses. Colton and Wiggins (2012) introduce the term *curation coefficient* as an informal reading of the *typicality*, *novelty* and *quality* measures put forward in (Ritchie 2007). In essence, this involves a project team member examining the output from their generative software, and calculating the proportion that they would be happy to present to others. For our purposes here, we used slightly lower criteria: we took all the ideas from each method, or a sample when there were too many, and recorded how many were suitable for assessment, i.e., the proportion of ideas that were both understandable and fictional, without any judgement of quality.

In figure 1, we presented flowcharts A to D for generating fictional ideas using ConceptNet. Facts in ConceptNet are scored for truth likelihood, and flowchart A is parametrised by a threshold, T_1, for the minimum score that ConceptNet facts must achieve to be used. Flowchart B uses ConceptNet twice, hence has thresholds T_1 and T_2. Flowcharts C and D were not parametrised, and used a fixed ConceptNet threshold of 1. Table 1 shows the number of ideas (yield) that each flowchart (FC) produced, with various threshold settings. The table also shows the curation coefficient (C-Coeff), i.e., the proportion of understandable and (largely) fictional ideas. We see that the yield reduces as higher thresholds T_1 and T_2 are imposed, but the curation coefficient increases, because fewer spurious or nonsensical facts are inverted for the ideas. In one case for flowchart B, by setting T_1 and T_2 to 5, we were able to produce a set of 27 ideas with a 100% curation coefficient. We noted an average yield of 190.4 and an average curation coefficient of 84.1%.

FC	Example	T_1	T_2	Yield	C-Coeff(%)
A	He was half man, half bird	1	-	97	72
		3	-	21	90
		5	-	14	93
B	He was half man, half fish, who could live in a lake	5	1	453	78
		5	2	94	88
		5	5	27	100
B	He was a cat, but he could still write	5	1	48	88
		5	3	7	100
C	Composer in a nest with turnip for a face	-	-	272	56
D	Dolphin that is made out of gold	-	-	871	76
	Average			190.4	84.1

Table 1: Curation analysis: ConceptNet approach.

Criteria	Yield	C-Coeff(%)
Fictional	500	90.9
Understandable	500	94.6
Non-duplicate	500	73.6
Overall	500	59.1

Table 2: Curation analysis: ReVerb approach.

Evaluation	Yield	C-Coeff(%)
What-if + moral (b-term)	32	28.1
What-if + moral (random)	16	6.25

Table 3: Curation analysis: bisociative discovery approach.

We generated 500 ideas with the ReVerb approach, using as seed queries the top six names from an online list of the most famous people of all time[6]. There were three issues with the ideas: (i) some happened to be true facts, or very close to a true fact (e.g., *What if John Kennedy was elected vice president?*); (ii) some happened to be nonsensical (e.g., *What if Elvis Presley is inducted into St?*), and (iii) some were an exact or very close duplicate of one already seen in the output (e.g., *What if Leonardo da Vinci was born in New York?* and *What if Leonardo was born in New York?*). In table 2, we report the curation coefficients with each of

[6] www.whoismorefamous.com

these three issues in mind, and an overall coefficient for the ideas which have none of these issues. We see that each issue reduced the curation coefficient, which was 59.1% overall.

For the bisociative discovery approach, we performed an analysis of the ideas that combine a What-if sentence with a moral statement, since these are automatically generated, rather than just mined from Twitter. We compared the 32 sentence pair ideas where there was a shared b-term with the 16 randomly concatenated pairs of sentences. Table 3 shows the results of the curation analysis for the ideas from the bisociative discovery approach. We found that the ideas generated by the bisociative discovery method were entirely understandable, as they were concatenations of two already understandable sentences. However, the results were often non-fictional, because the method doesn't explicitly attempt to distort reality. This explains the low curation curation coefficient of 28.1% for the b-term method, but it is important that it significantly outperformed the random approach.

With the ConceptNet and ReVerb approaches, data-mined notions of reality were inverted and altered respectively, hence the ideas were largely fictional. With respect to nonsensical ideas, for the ConceptNet-based ideas, we learned that control over quality could be exerted, at the expense of yield, through the usage of the ConceptNet thresholds. For the ReVerb results, completely nonsensical ideas were rare, since we used only arguments that are well attested with the relation. Errors were generally due to the open-domain IE extraction method used to compile the original facts. With the ReVerb approach, many of the (almost) true ideas occur because of substitutions for similar arguments, e.g., substituting 'president' with 'vice-president'. The system cannot recognise that the two are similar, and consequently the output contains a high proportion of almost exact duplicates: often almost the same thing is substituted many times over. This suggests that the results could be improved by incorporating a measure of semantic similarity which prefers dissimilar substitutions. Alternatively, the data integration technique from (Yao, Riedel, and McCallum 2012) could be used by the system to rule out ideas that, although not seen explicitly before, are highly probably repeats, given the observed facts.

A Crowd-Sourcing Evaluation

Ultimately, the fictional ideas we want to automatically produce will be for general consumption. Hence a large part of the WHIM project will involve crowd-sourcing responses to fictional ideas and using machine learning techniques to derive an audience model that can predict whether generated ideas are going to be of value. To study a baseline methodology for this, and to get a first tranche of feedback from the general public, we focused on the ConceptNet approach within the context of anthropomorphised animal characters which could feasibly appear in a Disney animated film. This context was chosen because Disney movies are familiar to most people and somewhat formulaic, hence we could be reasonably confident that when we surveyed people, our questions would be interpreted appropriately.

During a pilot study reported in (Llano et al. 2014), we focused on ideas generated by the CapableOf relation in the second ConceptNet node of flowchart B in figure 1, i.e., we studied ideas of the type: "What if there was a little X, who couldn't Y?" With an online survey of four questions, we asked 10 English speaking participants to rank the same list of 15 such Disney characters, in terms of (a) general impression (b) emotional response provoked (c) *narrative potential*: number and quality of potential plot lines imaginable for the character, and (d) how surprising they found the character to be. Our aim was to measure the influence of emotional provocation, narrative potential and surprise on general impression. Recall that we wrote routines to produce chains of ConceptNet facts. The 15 Disney characters in the survey comprised 5 from ideas with no chains, 5 from ideas with multiple chains, and 5 ideas where the RHS of a ConceptNet fact was replaced with a randomly chosen verb.

This pilot study showed that ConceptNet ideas were ranked much higher than the random ones for three questions, with average ranks of 5.21 vs. 10.98 for general impression, 6.08 vs. 11.5 for emotional provocation and 5.00 vs. 11.32 for potential for narrative potential. Within the ConceptNet examples, those with chains were ranked slightly higher than those without: average ranks of 4.78 vs. 5.21 for general impression, 3.42 vs. 6.08 for emotional response and 4.68 vs. 5.00 for narrative potential. However, when assessing levels of surprise, the random ideas were ranked as best with an average rank of 4.48 vs. 8.18 for ConceptNet ideas with no chains, and 8.44 for those with chains. On reflection, we determined that this resulted from an inconsistent interpretation of the word 'surprising'. We also found in the pilot study that there was a strong positive correlation r between general impression and both emotional response (r=0.81) and narrative potential (r=0.87), confirming that both these elements are key components of participants' general impressions of value. However, we found a strong negative correlation between general impression and surprise (r=-0.77). Hence, this suggests that more surprising ideas aren't generally well received.

Building on and learning from the pilot study, we undertook a larger scale experiment. For this, we used three sets of Disney characters generated using ConceptNet facts with the CapableOf (CO) relation as before, in addition to the Desires (D) relation ("What if there was a little X who was afraid of Y?") and the LocatedNear (LN) relation ("What if there was a little X who couldn't find the Y?") In order to evaluate participants' preferences, we designed four surveys: one per relation, and a fourth that mixed Disney characters from the three relations. In order to prevent bias or fatigue, each participant completed only one of the surveys.

Each survey consisted of four questions that asked participants to rank Disney characters in order of their general impression (GI) of the character's viability, the degree of emotional response (ER) they felt upon reading and interpreting the idea of the character, the quantity and quality of the plot lines; i.e., narrative potential (NP), that they felt might be written about each, and to what level each character met their expectation (LE) of a Disney character. This last question replaced the final question from the pilot study. The relation-focused surveys had a set of 14 ideas, eight ConceptNet non-chaining (NC) ideas (i.e., only one associated

Q	CO		D		LN		Avg	
	NC	CC	NC	CC	NC	CC	NC	CC
GI	7.41	7.62	7.76	7.15	8.05	6.77	7.74	7.18
ER	7.88	7.00	8.03	6.80	7.85	7.03	7.92	6.94
NP	7.85	7.04	8.03	6.80	7.95	6.90	7.94	6.91
LE	7.95	6.90	8.15	6.63	8.01	6.81	8.04	6.78

(a) Average participant rankings for three relation-focused surveys by type of idea: Non-Chaining (NC) and ConceptNet Chaining (CC).

Q	Mixed		
	CO	D	LN
GI	7.48	7.70	8.81
ER	6.55	8.44	9.01
NP	7.86	7.48	8.66
LE	7.24	8.46	8.30

(b) Average participant rankings for Mixed survey by inverted relation.

	GI&ER	GI&NP	GI&LE
Avg. Corr. (τ)	0.34	0.36	0.31

	ER&NP	ER&LE	NP&LE
Avg. Corr. (τ)	0.35	0.32	0.37

(c) Average rank correlation between all the questions of the four surveys: General Impression (GI), Emotional Response (ER), Narrative Potential (NP) and Level of Expectation (LE) .

Q	Correlation (τ)				
	CO	D	LN	Mixed	Avg
GI	0.09	0.25	0.27	-0.24	0.09
ER	0.17	0.25	0.26	0.26	0.23
NP	0.22	0.22	0.21	0.23	0.22
LE	0.14	0.27	0.22	0.08	0.17

(d) Rank correlation between av. participant rankings & chaining rankings.

Q	Correlation (τ)														
	CapableOf			Desires			LocatedNear			Mixed			Avg		
	IsA	CO	CB	IsA	D	CB	IsA	LN	CB	IsA	Rel	CB	IsA	Rel	CB
GI	0.25	0.19	0.31	0.42	0.17	0.40	-0.17	0.34	-0.17	0.20	0.27	0.31	0.17	0.24	0.21
ER	0.18	0.22	0.25	0.51	0.10	0.49	-0.07	0.21	-0.03	0.22	0.40	0.39	0.21	0.23	0.27
NP	-0.02	0.07	0.03	0.46	0.07	0.44	-0.07	0.27	-0.03	0.23	0.26	0.26	0.15	0.16	0.17
LE	0.39	0.11	0.44	0.46	0.10	0.44	0.02	0.17	0.06	0.18	0.29	0.31	0.26	0.16	0.31

(e) Rank correlation between average participant rankings and ConceptNet relations rankings.

Figure 2: Crowd-sourcing experiment results for four surveys: CapableOf (CO), Desires (D), LocatedNear (LN) and Mixed.

chain) and six ConceptNet chained (CC) ideas (i.e., with multiple associated chains) – random ideas were not evaluated as they scored significantly worse in the pilot study. The mixed-survey used a set of 15 CC-ideas, five per relation. These ideas were chosen by sampling systematically at equal intervals in terms of chaining score.

Results

A total of 135 participants completed the crowd sourcing experiment, with at least 27 participants per survey. Contrary to the pilot study, the crowd sourcing evaluation was not restricted to native English speakers. Therefore, we had respondents with different levels of fluency: 1 was at a basic level, 12 consider themselves at an intermediate level, 68 participants were fluent and 54 were native English speakers. These figures show that at least 90% of the participants were fluent or native, which provides a high level of confidence in the reliability of the results. Moreover, 64 participants were female, 70 were male and 1 person preferred not to specify their gender. This shows an almost even participation from both genders. The participants were between 18 and 74 years old; more specifically, 12 were in the age range between 18 and 24 years old, 74 in the range 25-34, 33 in the range 35-44, 7 in the range 45-54, 7 in the range 55-64 and 2 in the range 65-74. The highest concentration is seen in participants between 25 and 34 years old; however, most age ranges were represented in the surveys. After completing the surveys we asked the participants to select their level of confidence, between *very low, low, medium, high* and *very high*, when answering each question. Table 4 shows that most of the participants answered each question with a medium level of confidence or higher. This increases the confidence we have in the results.

Figure 2(a) shows the average rankings given for each class of ideas in the relation-focused surveys. As suggested in the pilot study, in general, the CC-ideas are ranked around

	Percentage of Participants			
Question	CO	D	LN	Mixed
GI	97	90	94	96
ER	97	90	88.5	92.5
NP	78	82.5	83	85
LE	85	80	80	78

Table 4: Percentage of participants who answered each question with a medium level of confidence or higher.

1 position higher than the NC-ideas. This supports the hypothesis that the ConceptNet chaining evaluation technique provides a reliable measure of value for fictional ideation using ConceptNet. Using a Friedman test comparing the mean ranks for CC and NC ideas in each response, we found that the difference between their ranks is highly significant overall ($p<0.001$). This effect remained significant across all question and survey subgroups.

Figure 2(b), which presents the results from the fourth survey, shows that, in general, the CO-ideas were ranked highest, followed by the D-ideas and then the LN-ideas. A Friedman test showed these differences to be highly significant overall ($p=0.001$). Our interpretation is that participants considered that, in some cases, the D-ideas and LN-ideas failed with respect to the feasibility of the fictional characters they portrayed, therefore, they were ranked lower. More specifically, respondents suggested that they felt apathy towards anthropomorphisations such as *'a little goat who is afraid of eating'* (D-idea), which threatened fundamental aspects of animals' lives, as well as ideas such as *'a little oyster who couldn't find the half shell'* (LN-idea), which were found difficult to interpret. On the contrary, participants pointed out that some of the CO-ideas were *"reminiscent of existing cartoons"*, placing them into a higher rank, e.g., *'a little bird who couldn't learn to fly'* (which resembles the plot of the animated film *Rio*). These type of participant

judgements played an important role when ranking the ideas, resulting in a clear overall preference for the CO-ideas.

We also wanted to confirm the pilot study suggestion that emotional response, narrative potential and level of expectation are key components of participants' general impression of value. We used a Kendall rank correlation coefficient (τ) for this analysis. Figure 2(c) shows the average correlation results between all the components, showing a positive correlation between all the surveyed components. However, a Friedman rank sum test indicated that the particular differences between correlation values are not significant (p=0.2438), i.e., all question pairs were similarly correlated.

Figure 2(d) shows the correlation between the chaining scores and the overall rankings of the participants. We see that weak positive correlations were found for most of the aspects evaluated in the four surveys and the chaining scores. These results confirm that, as suggested in the pilot study, the chaining technique can be used as a measure to evaluate fictional ideas, and we plan to investigate the value of generating other semantic chains to increase the effectiveness of this technique. Figure 2(d) also shows that a weak negative correlation exists between participants' general impression and the chaining scores for the mixed-survey. This suggests that participants found it more difficult to decide on the rankings when the rendering of the ideas was mixed.

Finally, two facts are used for each idea generated with ConceptNet: facts that tagged words as animals with the *IsA* relation, and facts to be inverted, which use the CapableOf, Desires and LocatedNear relations. Figure 2(e) shows the results of calculating the correlation between the average participants' rankings and each ConceptNet fact score, as well as the combination of both (CB). We see that, except for the LN-survey, most of the results show a weak positive correlation. This supports the finding from the pilot study that the values people project onto ideas is somewhat in line with the score assigned by ConceptNet to the underlying facts. Moreover, the highest correlations are presented in the D-survey with the *IsA* relation. We believe that people tend to rank higher ideas associated with more common animals, such as dogs or cats, used in multiple ideas of the D-survey, than ideas involving relatively uncommon animals, such as ponies, moles or oxen, which were used in the LN-survey.

The correlations between the participants' rankings and the chaining and ConceptNet scores (Figures 2(d) and 2(e)) led us to believe that these scores could be used to predict people's preferences when ranking fictional ideas. To test this hypothesis, we used the Weka machine learning framework (Hall et al. 2009). We provided Weka with the scores of: ConceptNet chaining, ConceptNet strength for the IsA relation, ConceptNet strength for the inverted relations, word frequencies for the LHS and RHS of inverted facts, and semantic similarity between the LHS and RHS of inverted facts, obtained using the DISCO system[7]. We classified each idea into *good* (top 5), *bad* (bottom 5) or *medium* (middle 5) based on the average participants' rankings. We tested a variety of decision tree, rule-based and other learning mechanisms, with the results given in Table 5, along with the name of the learning method which produced the best classifier. We found that the RandomTrees approach consistently performed well, but was only the best method for two aspects of evaluation. We used Weka to perform a Paired T-Test, which showed that the predictors are significantly better than the majority class classifier (MCC) – which simply assigns the largest class as a prediction – with up to 95% confidence.

[7] www.linguatools.de/disco/disco_en.html

	MCC	**GI**	**ER**	**NP**	**LE**
Method	ZeroR	Ridor	RandTree	NBTree	RandTree
Accuracy(%)	35.08	49.12	56.14	43.85	54.38

Table 5: Predictive accuracy for general impression, emotional response, narrative potential and level of expectation. Note that MCC value was the same for all evaluated aspects, i.e., GI, ER, NP and LE.

Conclusions and Future Work

While essential to the simulation of creative behaviour in software, fictional ideation has barely been studied in Computational Creativity research. Within the WHIM project, we have implemented three approaches to automated fictional ideation which act as a baseline to compare future ideation methods against. We presented baseline methodologies for assessment, in the form of a curation analysis and a crowd-sourcing study where participants ranked fictional ideas. The curation analysis showed that when guided in a strong context such as Disney characterisations, automated ideation methods work well, but they degrade when the context becomes weaker. The crowd sourcing study showed that an inference chaining technique – inspired by the hypothesis that ideas can be evaluated through narratives involving them – provides a reliable measure of value with which to assess the quality of fictional ideas. Also, we found positive correlations between the rankings of general impression and each of emotional response, narrative potential and expectation, showing that these are key elements of participants' general impression of fictional ideas. Finally, we demonstrated that machine learning techniques can be used to predict how people react to a fictional idea along these axes, albeit with only around 50% predictive accuracy.

The baselines presented here provide a firm foundation on which to build more intelligent ideation methods. We plan to improve open information extraction techniques for web mining, and to investigate ideation techniques involving metaphor and joke generation methods and the subversion of category expectations. Also, we plan to use extrapolation to explore scenarios that arise from a fictional idea. For instance, from the seed idea *What if there was an elevator with a million buttons?* we could extrapolate the distance the elevator can reach and come up with a scenario in which elevators can reach as high as space. Identifying that the current distance reached by elevators is significantly lower than the distance to space is crucial in order to select this idea as an interesting scenario. Using quantitative information can help achieve this goal. The Visuo system (Gagné and Davies 2013) uses semantic similarity to estimate quantitative information for input descriptions of scenes by transferring quantitative knowledge to concepts from distributions of familiar

concepts in memory. We will explore the use of Visuo in the production of scenarios from a fictional idea.

The generation and assessment of narratives will be a key factor, enabling the system to curate its output. We will derive a theory of idea-centric narratives and implement methods for generating them and assessing ideas in terms of the quality/quantity of narratives they appear in. Our ConceptNet chaining technique shows much promise. Based on the correlation found between general impression and emotional response, we plan to improve the predictive power of the technique using sentiment analysis, as in (Liu, Lieberman, and Selker 2003), where the affect of a concept is assessed through a chaining process. The final major aspects will be to experiment with rendering methods where obfuscation and affect are used to increase audience appreciation of an idea; and the machine learning of a detailed audience model which will influence the entire ideation process.

The WHIM project is primarily an engineering effort to build a What-if Machine as a web service and interactive engine, which generates fictional ideas, and provides motivations and consequences for each idea, potential narratives involving it, and related renderings such as poems, jokes, neologisms and short stories. The first version of the What-if Machine is available online[8], and uses Flowchart E from figure 1. Users can parametrise the method for exploration, or simply click the *'I'm feeling lucky'* button. This online implementation will be used to gather feedback for audience modelling, and hopefully help promote fictional ideation as a major new area for Computational Creativity research.

Acknowledgements

We would like to thank the members of the Computational Creativity Group at Goldsmiths for their feedback, Jasmina Smailović for preprocessing the tweets used in the bisociative approach, the participants of the crowd sourcing study for their time, and the anonymous reviewers for their constructive comments. This research was funded by the Slovene Research Agency and supported through EC funding for the project WHIM 611560 by FP7, the ICT theme, and the Future Emerging Technologies FET programme.

References

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the International Conference on Management of Data*.

Charnley, J.; Colton, S.; and Llano, M. T. 2014. The FloWr framework: Automated flowchart construction, optimisation and alteration for creative systems. In *Proceedings of the 5th International Conference on Computational Creativity*.

Colton, S., and Wiggins, G. 2012. Computational Creativity: The final frontier? In *Proceedings of the 20th European Conference on Aritificial Intelligence*.

Fader, A.; Soderland, S.; and Etzioni, O. 2011. Identifying relations for open information extraction. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Fauconnier, G., and Turner, M. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Gagné, J., and Davies, J. 2013. Visuo: A model of visuospatial instantiation of quantitative magnitudes. *The Knowledge Engineering Review* 28(3):347–366.

Goel, A. K. 2013. Biologically inspired design: A new program for computational sustainability. *IEEE Intelligent Systems* 28(3):80–84.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10–18.

Juršič, M.; Cestnik, B.; Urbančič, T.; and Lavrač, N. 2012. Cross-domain literature mining: Finding bridging concepts with crossbee. In *Proceedings of the 3rd International Conference on Computational Creativity*.

Koestler, A. 1964. *The act of creation*, volume 13. Hutchinson & Co.

Lin, T.; Mausam; and Etzioni, O. 2012. Entity linking at web scale. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*.

Liu, H., and Singh, P. 2004. Commonsense reasoning in and over natural language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*. Springer.

Liu, H.; Lieberman, H.; and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*.

Llano, M.; Hepworth, R.; Colton, S.; Charnley, J.; and Gow, J. 2014. Automating fictional ideation using ConceptNet. In *Proceedings of the AISB14 Symposium on Computational Creativity*.

Martins, J.; Pereira, F.; Miranda, E.; and Cardoso, A. 2004. Enhancing sound design with conceptual blending of sound descriptors. In *Proceedings of the 1st Joint Workshop on Computational Creativity*.

Moorman, K., and Ram, A. 1996. The role of ontology in creative understanding. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*.

Pereira, F., and Cardoso, A. 2003. The horse-bird creature generation experiment. *The Interdisciplinary Journal of Artificial Intelligence and the Simulation of Behavior* 1(3):257–280.

Pereira, F., and Gervás, P. 2003. Natural language generation from concept blends. In *Proceedings of the AISB Symposium on AI and Creativity in Arts and Science*.

Pereira, F. 2007. *Creativity and AI: A Conceptual Blending Approach*. Mouton de Gruyter.

Perovšek, M.; Cestnik, B.; Urbančič, T.; Colton, S.; and Lavrač, N. 2013. Towards narrative ideation via cross-context link discovery using banded matrices. In *Proceedings of Advances in Intelligent Data Analysis XII*.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:67–99.

Veale, T. 2006. Tracking the lexical zeitgeist with WordNet and Wikipedia. In *Proceedings of 17th European Conference on Artificial Intelligence*.

Yao, L.; Riedel, S.; and McCallum, A. 2012. Probabilistic databases of universal schema. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*.

[8] www.whim-project.eu/whatifmachine