Weaving creativity into the Semantic Web: a language-processing approach

Article (Published Version)

Weaving creativity into the Semantic Web: a language-processing approach

Anna Jordanous
Centre for e-Research
Department of Digital Humanities
King's College London, UK
anna . jordanous at kcl . ac . uk

Bill Keller
Department of Informatics
University of Sussex
Brighton, UK
billk at sussex . ac . uk

Abstract

This paper describes a novel language processing approach to the analysis of creativity and the development of a machine-readable ontology of creativity. The ontology provides a conceptualisation of creativity in terms of a set of fourteen key components or *building blocks* and has application to research into the nature of creativity in general and to the evaluation of creative practice, in particular. We further argue that the provision of a machine readable conceptualisation of creativity provides a small, but important step towards addressing the problem of automated evaluation, 'the Achilles' heel of AI research on creativity' (Boden 1999).

Introduction

Creativity is a complex, multi-faceted concept encompassing many related aspects, abilities, properties and behaviours. This complexity makes the production of a comprehensive and generally applicable account of creativity problematic. Existing definitions of creativity are often too superficial for use by the research community and may be subject to discipline or domain bias, limiting their application. The need for a comprehensive, multi-dimensional account has been widely recognised (Rhodes 1961; Torrance 1967; Plucker, Beghetto, and Dow 2004; Kaufman 2009). Such an account would assist our understanding of creativity, highlighting areas of common ground and avoiding the pitfalls of disciplinary bias (Hennessey and Amabile 2010; Plucker and Beghetto 2004).

Words associated with academic debate about the nature of creativity are strongly linked to our understanding of its meaning and attributes. Analysis of this language provides a sound basis for constructing a sufficiently detailed and comprehensive account of the concept. In the present work, statistical language processing techniques are used to identify words significantly associated with creativity in a corpus of academic papers on the topic. A measure of lexical similarity provides a basis for clustering words and identifying key themes or components of creativity. The set of components yields information about the nature of creativity, based on what we emphasise when we discuss the concept.

Within the field of computational creativity, the problem of automatic evaluation remains a significant issue: 'the Achilles' heel of AI research on creativity' (Boden 1999). Recently, the Semantic Web has emerged as a way to address the troublesome but important issue (Boden 1999) of articulating values, concepts and information in an open and *machine-readable* format. Linked Data is the term used in the Semantic Web community to describe published data that is machine-readable and connected together using semantically typed links. We take the step of encoding our components in RDF, the current W3C standard for implementing Linked Data.[1] The resulting ontology is available to the wider research community as a resource in the Semantic Web, under the permanent URI `http://purl.org/creativity/ontology`, a form familiar to Semantic Web researchers and also accessible through browsers such as Marbles.[2]

Currently, most content on the Semantic Web is in the form of ontologies of 'things': semantically structured collections of factual or objective data on topics as diverse as people, places, narratives, or music.[3] To date, little work has been done on specifically defining subjective concepts in an ontology. However, current work on lexical resources such as WordNet has laid foundations for more definitionally troublesome concepts to be considered in detail; the time is ripe for development of ontologies of subjective concepts such as creativity.

Components of creativity

We identify a core lexicon consisting of just those words that appear to be highly associated with discussions of creativity in a corpus of academic papers on the topic. Our approach substantially develops and refines work described in Jordanous (2010). A key innovation is the use of a measure of lexical similarity, which allows the words to be clustered automatically to reveal a number of common themes or factors of creativity. Further analysis results in a set of fourteen

[1] `http://www.w3.org/TR/rdf-syntax-grammar`, last accessed 27th January 2012.

[2] `http://www.w3.org/2001/sw/wiki/Marbles`, last accessed 27th January 2012.

[3] Example ontologies are available at `http://www.foaf-project.org`, `http://www.geonames.org/ontology`, `http://www.contextus.net/ontomedia` and `http://musicontology.com` respectively, all last accessed 27th January 2012.

key components.

Corpus data

A 'creativity corpus' was assembled from a sample of 30 academic papers examining creativity from a variety of stand-points (Jordanous 2010). The selected papers cover a wide range of years (1950-2009) and academic disciplines, from psychological studies to computational models. Academic papers were used due to ease of location (e.g. through targeted literature search), accessibility (electronic publication for download), format (ease of conversion to text allows for computational analysis) and availability of citation data (used as a criterion for inclusion of a paper).[4]

In Jordanous (2010), language use in the creativity corpus was compared to general language use as represented by the British National Corpus (BNC) (Leech 1992). This had the undesired effect of highlighting words that were predominant in academic papers but not necessarily specific to creativity literature, e.g. *et', al'*. In the present study, a further corpus of 60 academic papers on topics unrelated to creativity was assembled (a 'non-creativity corpus'). For each paper in the creativity corpus, we retrieved the two most-cited papers in the same academic discipline[5] and with the same year of publication, that did not contain any words with the prefix *creat* (i.e. *creativity*, *creative*, *creation*, etc.).

Each corpus was processed using the RASP natural language processing toolkit (Briscoe, Carroll, and Watson 2006) to perform lemmatisation and part-of-speech (POS) tagging. Lemmatisation allows us to ignore morphological variation so that, e.g., *processed* and *processing* are both recognised as forms of *process*. POS tagging allows us to distinguish between different grammatical usages of the same orthographical form: e.g. *process* as a noun or as a verb. Two lists of frequency counts were produced: one for all words occurring in the creativity corpus and one for all words in the non-creativity corpus. Only 'content-bearing' words (i.e. nouns, verbs, adjectives and adverbs) were considered to be of interest. Any 'function words' or other minor categories (pronouns, articles, prepositions etc.), were ignored as they have little or no independent semantic content and are therefore of limited interest for the present study.

Finding words associated with creativity

A standard, statistical measure of association was used to identify words salient to discussions of creativity. The log-likelihood ratio (or G-squared statistic) is a measure of how well observed frequency data fit a model or expected frequency distribution. The statistic is an alternative to Pearson's chi-squared (χ^2) test that has been advocated as a more appropriate measure for corpus analysis as it does not rely on the (unjustifiable) assumption of normality in word distribution (Dunning 1993). This is a particular issue when analysing relatively small corpora as in the present case.[6] The log likelihood ratio is more accurate than χ^2 in its treatment of infrequent words in the data, which often hold useful information.

Our use of the log-likelihood ratio follows that of Rayson and Garside (2000). Given two corpora (in our case, 'creativity corpus' and 'non-creativity corpus') the log-likelihood score for a given word is calculated as:

$$LL = 2 \sum_{i \in \{1,2\}} O_i \ln(\frac{O_i}{E_i}) \quad (1)$$

where O_i is the observed frequency of the given word in corpus i and E_i is its expected frequency in corpus i. The expected frequency E_i is given by:

$$E_i = \frac{N_i \times (O_1 + O_2)}{N_1 + N_2} \quad (2)$$

where N_i denotes the total number of words in corpus i.

Following standard statistical practice, any word occurring fewer than five times was excluded. This ensures that the statistics are robust. To identify significant results, we also removed words with a log-likelihood score less than 10.83, representing a chi-squared significance value for p=0.001 (one degree of freedom). To identify words strongly associated with discussion of creativity it was necessary to select just those words with observed counts higher than than expected in the creativity corpus. This resulted in a total of 694 distinctive *creativity words*: a collection of 389 nouns, 205 adjectives, 72 verbs and 28 adverbs that occurred significantly more often than expected in the creativity corpus. The 20 such words with the highest log-likelihood ratio scores are listed in Table 1.

It is important to note that our objective is to identify key themes in the lexical data, not to induce a comprehensive terminology of creativity. Despite the relatively small size of the available corpora, the resulting set of 694 creativity words is sufficiently rich for this purpose.

Identifying components of creativity

In Jordanous (2010) an attempt was made to identify key components by clustering creativity words by inspection of the raw data. In practice, this proved laborious and made it impossible systematically to consider all of the identified words. It also raised issues of subjectivity and experimenter bias. Here we address these problems, at least in part, by first clustering all the words automatically according to a statistical measure of *distributional similarity* (Lin 1998). The more manageable collection of clusters are then inspected manually to identify key components.

Intuitively, words that tend to occur in similar linguistic contexts will tend to be similar in meaning (Harris 1968). For example, evidence that the words *concept* (LLR=189.90) and *idea* (LLR=475.74) are similar in meaning might be provided by occurrences such as the following:

[4] Note that some papers have been published in very recent years and therefore have few citations. In this case selection was based on subjective judgement of influence.

[5] As categorised by the literature database *Scopus* `http://www.scopus.com/`), last accessed 27th January 2012.

[6] At around 300K and 700K words respectively, the creativity and non-creativity corpora are very small compared to the British National Corpus ($\approx$ 100M words) and tiny in comparison to recent, web-derived text collections of billions of words.

Word (and part of speech tag)	LLR
thinking (N)	834.55
process (N)	612.05
innovation (N)	546.20
idea (N)	475.74
program (N)	474.41
domain (N)	436.58
cognitive (J)	393.79
divergent (J)	355.11
openness (N)	328.57
discovery (N)	327.38
primary (J)	326.65
originality (N)	315.60
criterion (N)	312.61
intelligence (N)	309.31
ability (N)	299.27
knowledge (N)	290.48
create (V)	280.06
experiment (N)	253.32
plan (N)	246.29
agent (N)	246.24

Table 1: The top 20 results of the log-likelihood ratio (LLR) calculations. A significant LLR score at p=0.001 is 10.83.

1. the *concept/idea* involves (subject of verb 'involve')
2. applied the *concept/idea* (object of verb 'apply')
3. the basic *concept/idea* (modified by adjective 'basic')

Word occurrence data of this kind was obtained from an analysis of the written portion of the BNC, which had previously been processed using the RASP toolkit to extract grammatical dependency relations (*subj-of*, *obj-of*, *modified-by*). Each word in the creativity corpus was then associated with a list of all of the grammatical relations in which it participated, together with corresponding counts of occurrence.

Distributional similarity of two words is measured in terms of the similarity of their associated lists of grammatical relations. The present work adopts an information-theoretic measure devised by Lin (1998), which has been widely used in language processing applications and shown to perform well against other similarity measures as a means of identifying near-synonyms (Weeds and Weir 2003) . Similarity scores were obtained separately for pairs of nouns, pairs of verbs and so on. For a given set of words, the similarity data is conveniently visualised as a graph or network, where nodes correspond to words and edges are weighted by similarity scores, as in Figure 1.

A possible problem with obtaining word similarity data this way would arise if the majority of the creativity words were used with distinctive or technical senses within the creativity corpus. This is unlikely, however: whilst some narrowly specialised usage may be present in our creativity lexicon, most words retain general senses reflected in the wider BNC data set.

The graph clustering software *Chinese Whispers* (Biemann 2006) was used to automatically identify word clusters in the dataset. This algorithm uses an iterative process to group together graph nodes that are located 'close' to each other. By grouping words with similar meanings, the number of data items was effectively reduced and themes in the data could be identified more readily by inspection. Themes discovered through clustering were further analysed in terms of the *Four Ps* of creativity (Rhodes 1961; Mooney 1963; MacKinnon 1970) to identify alternative perspectives and reveal subtler (but still important) aspects of creativity. From the analysis it was possible to extract a set of fourteen key components of creativity.

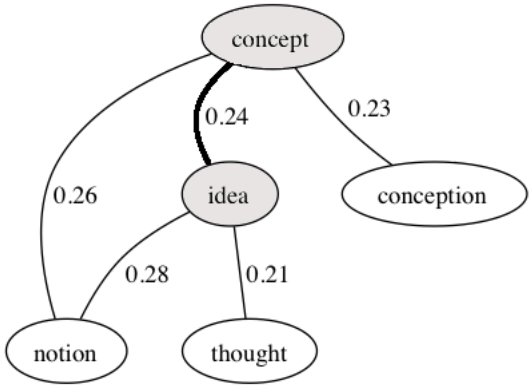

Figure 1: Graph representation of the similarity of the nouns *concept* and *idea* and related words. Words are drawn as nodes linked by weighted edges representing word similarity (maximum similarity is 1.0).

Implementing an ontology of creativity

The fourteen components provide a clear account of the constituent parts of the concept of creativity. Our remaining contribution is to express these components in a machine-readable form. We also want to use Linked Data principles (Heath and Bizer 2011) to connect the individual components to other data sources within the Semantic Web, so that creativity is defined in terms of concepts that have already been defined. To achieve this, we used SKOS (Simple Knowledge Organisation System),[7] a W3C standard which provides a model for representing ontological data within the Semantic Web. We also made use of WordNet (Reed and Lenat 2002), a large lexical database of English in which words are grouped by sense and interlinked by lexical and conceptual relations. WordNet has recently been made available as a Semantic Web ontology.[8]

The SKOS ontology incorporates three main classes: *skos:Concept* (anything we may want to record information about), *skos:ConceptScheme* (a set that collectively defines a skos:Concept) and *skos:Collection* (a collection of semantically-related information).

We created an instance of *skos:ConceptScheme* called

[7] `http://www.w3.org/TR/skos-reference`, last accessed 27th January 2012.

[8] `http://wordnet.rkbexplorer.com/`

CreativityComponents to represent the set of components that defines the *skos:Concept* of *Creativity*. Each component is represented as an individual *skos:Concept*. As RDF is a graph-based model, the resulting encoding can be visualised as in Figure 2. The graph has also been published in serialised format as an RDF/XML text file and made available as `http://purl.org/creativity/ontology`. The *skos:Concept* labelled *Creativity* has the unique URI `purl.org/creativity/ontology#Creativity` and any Linked Data that needs to refer to the concept can use this identifier.

The distributed nature of Semantic Web research means that the enormous task of defining concepts in a machine-readable form is divided across the research field, rather than being the sole responsibility of one particular research group. This work practice acts as a form of peer review, as ontologies are developed, critiqued, and ultimately judged by the extent to which they are adopted and re-used as points of reference by other researchers.

Upper ontologies allow us to link the concepts in our ontology to related ontological work on creativity in the future (even if these future researchers are not aware of our ontological contribution). An upper ontology defines higher-level vocabularies and concepts necessary to implement ontologies themselves, providing the meta-vocabulary to link specific ontologies to more general concepts. The implementation of the Wordnet dataset and structure as an ontology provides WordNet as an upper ontology for us to use, linking a lexical string (e.g. "creativity") to various concepts associated with that string, such as its sense, hyponyms, type, 'gloss' (brief definition) and other related lexical information.

Each component in our ontology is comprised of a cluster of keywords. It makes sense, therefore, to link each component back to the appropriate keywords, using the WordNet ontology at `http://wordnet.rkbexplorer.com/`. In this way, our components are linked into the Semantic Web through the WordNet ontology. This linkage also provides further semantic information on each component via the lexical relations and other information represented in the WordNet hierarchy. Finally, following Linked Data principles, we also link our interpretation of creativity as an extension of the representation of the concept in WordNet. In this way, machines (and people) can see the relationship between this general concept of creativity and our more detailed ontological analysis.

Discussion and Implications

The current work is part of a wider project engaged with the question of the evaluation of computational creativity (Jordanous 2011). The components of creativity have already been applied, both for in-depth expert evaluation and in forming snapshot judgements of the creativeness of a given system. The resulting component-based evaluation yields detailed information about creative strengths and weaknesses. Crucially, the evaluation highlights those components where a system performs poorly, providing insight into areas where improvement in performance is needed.

By publishing the ontology in the Semantic Web we ensure that it is freely available to the research community. This has a number of implications. First, it may be freely referred to, extended or amended. Refinement is clearly possible, for example in providing more fine-grained analysis of the components or in articulating the relationships between them. Second, it facilitates the development of creativity-aware applications to support manual evaluation of creativity based on the components. It also represents a step towards the development of methods of automated evaluation. One intriguing possibility is to further exploit language processing techniques to provide automated evaluation by proxy based on textual reviews or descriptions of system performance. This is analogous to the way that sentiment analysis techniques are now used to automatically evaluate attitude and opinion based on reviews of products or services (Pang and Lee 2008).

The current work illuminates the sorts of issues that arise in formal modelling of subjective or 'soft' concepts such as creativity. For example, some of our components appear logically inconsistent with others in the set: e.g. the need for autonomous, independent behaviour (*Independence and Freedom*) versus the requirement for social interaction (*Social Interaction and Communication*). Also, creativity clearly manifests itself in different ways across different domains (Plucker and Beghetto 2004) and components will vary in importance, according to the requirements of a particular domain. For example, creative behaviour in mathematical reasoning has more focus on finding a correct solution to a problem than is the case for creative behaviour in, say, musical improvisation (Colton 2008). Questions remain about how such dialectical and fluid aspects might be modelled. We present the set of components as a rather loose collection of dimensions – attributes, abilities and behaviours, etc. – which contribute to our overall understanding of creativity, rather than a unified definition.

Concluding remarks

This paper has described the development of an ontology of creativity using corpus-based, language processing techniques and its publication as machine-readable, Linked Data in the Semantic Web. The resulting ontology provides a multi-perspective analysis of creativity in terms of a set of fourteen key components and has application to the study and evaluation of computational creativity. Weaving the ontology into the Semantic Web has implications for future work on modelling subjective concepts and suggests some interesting directions for future research into the problem of automated evaluation of creativity.

References

Biemann, C. 2006. Chinese Whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing*, 73–80. Morristown, NJ: Association for Computational Linguistics.

Boden, M. A. 1999. Introduction [summary of Boden's keynote address to AISB'99]. In *AISB Quarterly - Special issue on AISB99: Creativity in the arts and sciences*, volume 102, 11.

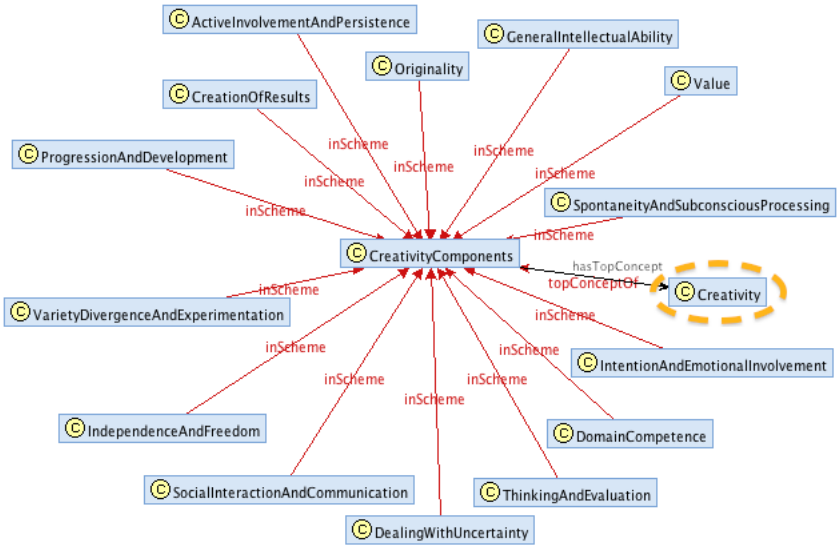

Figure 2: The RDF ontology of Creativity, in graph form.

Briscoe, E.; Carroll, J.; and Watson, R. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *Proceedings of AAAI Symposium on Creative Systems*, 14–20.

Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1):61–74.

Harris, Z. 1968. *Mathematical Structures of Language*. New York: Wiley.

Heath, T., and Bizer, C. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool.

Hennessey, B. A., and Amabile, T. M. 2010. Creativity. *Annual Review of Psychology* 61:569–598.

Jordanous, A. 2010. Defining creativity: Finding keywords for creativity using corpus linguistics techniques. In *Proceedings of the International Conference on Computational Creativity*, 278–287.

Jordanous, A. 2011. Evaluating evaluation: Assessing progress in computational creativity research. In *Proceedings of the Second International Conference on Computational Creativity (ICCC-11)*.

Kaufman, J. C. 2009. *Creativity 101*. The Psych 101 series. New York: Springer.

Leech, G. 1992. 100 million words of English: the British National Corpus (BNC). *Language Research* 28(1):1–13.

Lin, D. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, 296–304.

MacKinnon, D. W. 1970. Creativity: a multi-faceted phenomenon. In Roslansky, J. D., ed., *Creativity: A Discussion at the Nobel Conference*. Amsterdam, The Netherlands: North-Holland Publishing Company. 17–32.

Mooney, R. L. 1963. A conceptual model for integrating four approaches to the identification of creative talent. In Taylor, C. W., and Barron, F., eds., *Scientific Creativity: Its Recognition and Development*. New York: John Wiley & Sons. chapter 27, 331–340.

Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval* 2(1-2):1–135.

Plucker, J. A., and Beghetto, R. A. 2004. Why creativity is domain general, why it looks domain specific, and why the distinction doesn't matter. In Sternberg, R. J.; Grigorenko, E. L.; and Singer, J. L., eds., *Creativity: From Potential to Realization*. Washington, DC: American Psychological Association. chapter 9, 153–167.

Plucker, J. A.; Beghetto, R. A.; and Dow, G. T. 2004. Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist* 39(2):83–96.

Rayson, P., and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proceedings of ACL Workshop on Comparing Corpora*.

Reed, S. L., and Lenat, D. B. 2002. Mapping ontologies into Cyc. In *Proceedings of AAAI'02 workshop on Ontologies and the Semantic Web*.

Rhodes, M. 1961. An analysis of creativity. *Phi Delta Kappan* 42(7):305–310.

Torrance, E. P. 1967. Scientific views of creativity and factors affecting its growth. In Kagan, J., ed., *Creativity and Learning*. Boston: Beacon Press. 73–91.

Weeds, J., and Weir, D. 2003. Finding and evaluating nearest neighbours. In *Proceedings of the 2nd Conference of Corpus Linguistics*.