Creating Similarity: Lateral Thinking for Vertical Similarity Judgments

Tony Veale
Web Science and Technology Division,
Korean Advanced Institute of Science
and Technology, Yuseong, South Korea
Tony.Veale@gmail.com

Guofu Li
School of Computer Science and Informatics,
University College Dublin,
Belfield, Dublin D2, Ireland.
li.guofu.l@gmail.com

Abstract

Just as *observing* is more than just *seeing*, *comparing* is far more than mere *matching*. It takes understanding, and even inventiveness, to discern a useful basis for judging two ideas as similar in a particular context, especially when our perspective is shaped by an act of linguistic creativity such as metaphor, simile or analogy. Structured resources such as WordNet offer a convenient hierarchical means for converging on a common ground for comparison, but offer little support for the divergent thinking that is needed to creatively view one concept as another. We describe such a means here, by showing how the web can be used to harvest many divergent views for many familiar ideas. These lateral views complement the narrow vertical view offered by WordNet, and support a system for creative idea exploration called *Thesaurus Rex*. We also show how *Thesaurus Rex* supports a novel, generative similarity measure for WordNet.

1 Seeing is Believing (*and* Creating)

Similarity is a cognitive phenomenon that is both complex and subjective, yet for practical reasons it is often modeled as if it were simple and objective. This makes sense for the many situations where we want to align our similarity judgments with those of others, and thus focus on the same conventional properties that others are also likely to focus upon. This reliance on the consensus viewpoint explains why WordNet (Fellbaum, 1998) has proven so useful as a basis for computational measures of lexico-semantic similarity (e.g. see Pederson *et al*. 2004, Budanitsky & Hirst, 2006; Seco *et al.* 2006). These measures reduce the similarity of two lexical concepts to a single number, by viewing similarity as an objective estimate of the overlap in their salient qualities. This convenient perspective is poorly suited to comparisons that are creative or insightful, yet it is sufficient for the many mundane comparisons that one tacitly performs in daily life, such as when we organize our books or look for items in a supermarket. So if we do not know in which aisle to locate a given item (such as *oatmeal*), we may tacitly know how to locate a similar product (such as *cornflakes*) and orient ourselves accordingly.

Yet there are occasions when the recognition of similarities spurs the *creation* of similarities, when the act of comparison spurs us to invent new ways of looking at an idea. By placing *pop tarts* in the breakfast aisle, food manufacturers encourage us to view them as a breakfast food that is not dissimilar to *oatmeal* or *cornflakes*. When ex-PM Tony Blair published his memoirs, a mischievous activist encouraged others to move his book from *Biography* to *Fiction* in bookshops, in the hope that buyers would see it in a new light. Whenever we use a novel metaphor to convey a non-obvious viewpoint on a topic, such as "*cigarettes are time bombs*", the comparison spurs an audience to insight, to see aspects of the topic that *make* it more similar to the vehicle (see Ortony, 1979; Veale & Hao, 2007).

In formal terms, assume agent A has an insight about concept X, and uses the metaphor *X is a Y* to also provoke this insight in agent B. To arrive at

this insight for itself, B must intuit what X and Y have in common. But this commonality is surely more than a standard categorization of X, or else it would not count as an insight about X. To understand the metaphor, B must place X in a new category, so that X can be seen as more similar to Y. Metaphors shape the way we perceive the world by re-shaping the way we make similarity judgments. So if we want to imbue computers with the ability to make and to understand creative metaphors, we must first give them the ability to look beyond the narrow viewpoints of conventional resources.

Any measure that models similarity as an objective function of a conventional worldview employs a *convergent* thought process. Using WordNet, for instance, a similarity measure can vertically converge on a common superordinate category of both inputs, and generate a single numeric result based on their distance to, and the information content of, this common generalization. So to find the most conventional ways of seeing a lexical concept, one simply ascends a narrowing concept hierarchy, using a process de Bono (1970) calls *vertical thinking*. To find novel, non-obvious and useful ways of looking at a lexical concept, one must use what Guilford (1967) calls *divergent thinking* and what de Bono calls *lateral thinking*. These processes cut across familiar category boundaries, to simultaneously place a concept in many different categories so that we can see it in many different ways.

de Bono argues that vertical thinking is selective while lateral thinking is generative. Whereas vertical thinking concerns itself with the "right" way or a single "best" way of looking at things, lateral thinking focuses on producing alternatives to the status quo. To be as useful for creative tasks as they are for conventional tasks, we need to re-imagine our computational similarity measures as generative rather than selective, expansive rather than reductive, divergent as well as convergent and lateral as well as vertical. Though WordNet is ideally structured to support vertical, convergent reasoning, its comprehensive nature means it can also be used as a solid foundation for building a more lateral and divergent model of similarity. Here we will use the web as a source of diverse perspectives on familiar ideas, to complement the conventional and often narrow views codified by WordNet.

Section 2 provides a brief overview of past work in the area of similarity measurement, before section 3 describes a simple bootstrapping loop for acquiring richly diverse perspectives from the web for a wide variety of familiar ideas. These perspectives are used to enhance a WordNet-based measure of lexico-semantic similarity in section 4, by broadening the range of informative viewpoints the measure can select from. Similarity is thus modeled as a process that is both generative *and* selective. This lateral-and-vertical approach is evaluated in section 5, on the Miller & Charles (1991) dataset. A web app for the lateral exploration of diverse viewpoints, named *Thesaurus Rex*, is also presented, before closing remarks are offered in section 6.

2 Related Work and Ideas

WordNet's taxonomic organization of noun-senses and verb-senses – in which very general categories are successively divided into increasingly informative sub-categories or instance-level ideas – allows us to gauge the overlap in information content, and thus of meaning, of two lexical concepts. We need only identify the deepest point in the taxonomy at which this content starts to diverge. This point of divergence is often called the LCS, or *least common subsumer*, of two concepts (Pederson *et al*., 2004). Since sub-categories add new properties to those they inherit from their parents – Aristotle called these properties the *differentia* that stop a category system from trivially collapsing into itself – the depth of a lexical concept in the taxonomy is an intuitive proxy for its information content. Wu & Palmer (1994) thus use the depth of a lexical concept in the WordNet hierarchy as a proxy for its information content, and estimate the similarity of two lexical concepts as twice the depth of their LCS divided by the sum of their individual depths.

Leacock and Chodorow (1998) instead use the length of the shortest path between two concepts as a proxy for the conceptual distance between them. To connect two ideas in a hierarchical system, one must vertically ascend the hierarchy from one concept, change direction at a potential LCS, and then descend the hierarchy to reach the second concept. (Aristotle was also first to suggest this approach in his *Poetics*). Leacock and Chodorow normalize the length of this path by dividing its size (in nodes) by twice the depth of the deepest concept in the hierarchy; the latter is an upper bound on the distance between any two concepts in the hierarchy. Negating the log of this normalized length yields a corresponding similarity score. While the role of an

LCS is merely implied by Leacock and Chodorow's hierarchical use of a *shortest path,* the LCS is pivotal nonetheless, and like that of Wu & Palmer, the approach uses an essentially vertical reasoning process to identify a single "best" generalization.

Depth is a convenient proxy for information content, but more nuanced proxies can yield more rounded similarity measures. Resnick (1995) draws on information theory to define the information content of a lexical concept as the negative log likelihood of its occurrence in a corpus, either explicitly (via a direct mention) or by presupposition (via a mention of any of its sub-categories or instances). Since the likelihood of a general category occurring in a corpus is higher than that of any of its sub-categories or instances, such categories are more predictable, and less informative, than rarer categories whose occurrences are less predictable and thus more informative. The negative log likelihood of the most informative LCS of two lexical concepts offers a reliable estimate of the amount of information shared by those concepts, and thus a good estimate of their similarity. Lin (1998) combines the intuitions behind Resnick's metric and that of Wu and Palmer to estimate the similarity of two lexical concepts as an information ratio: twice the information content of their LCS divided by the sum of their individual information contents.

Jiang and Conrath (1997) consider the converse notion of *dissimilarity*, noting that two lexical concepts are dissimilar to the extent that each contains information that is not shared by the other. So if the information content of their most informative LCS is a good measure of what they *do* share, then the sum of their individual information contents, minus twice the content of their most informative LCS, is a reliable estimate of their dissimilarity.

Seco *et al.* (2006) presents a minor innovation, showing how Resnick's notion of information content can be calculated without the use of an external corpus. Rather, when using Resnick's metric (or that of Lin, or Jiang and Conrath) for measuring the similarity of lexical concepts in WordNet, one can use the category structure of WordNet itself to estimate information content. Typically, the more general a concept, the more descendants it will possess. Seco *et al*. thus estimate the information content of a lexical concept as the log of the sum of all its unique descendants (both direct and indirect), divided by the log of the total number of concepts in the entire hierarchy. Not only is this *intrinsic* view of information content convenient to use, without recourse to an external corpus, Seco *et al*. show that it offers a better estimate of information content than its extrinsic, corpus-based alternatives, as measured relative to the average similarity ratings offered by humans for the 30 word-pairs in the Miller & Charles (1991) test set.

A similarity measure can draw on other sources of information besides WordNet's category structures. One might eke out additional information from WordNet's textual glosses, as in Lesk (1986), or use category structures other than those offered by WordNet. Looking beyond WordNet, entries in the online encyclopedia Wikipedia are not only connected by a dense topology of lateral links, they are also organized by a rich hierarchy of overlapping categories. Strube and Ponzetto (2006) show how Wikipedia can support a measure of similarity (and relatedness) that better approximates human judgments than many WordNet-based measures. Nonetheless, WordNet can be a valuable component of a hybrid measure, and Agirre *et al.* (2009) use an SVM (support vector machine) to combine information from WordNet with information harvested from the web. Their best similarity measure achieves a remarkable **0.93** correlation with human judgments on the Miller & Charles word-pair set.

Similarity is not always applied to pairs of concepts; it is sometimes analogically applied to pairs *of pairs* of concepts, as in proportional analogies of the form *A is to B as C is to D* (e.g., *hacks are to writers as mercenaries are to soldiers*, or *chisels are to sculptors as scalpels are to surgeons*). In such analogies, one is really assessing the similarity of the unstated relationship between each pair of concepts: thus, mercenaries are soldiers whose allegiance is paid for, much as hacks are writers with income-driven loyalties; sculptors use chisels to carve stone, while surgeons use scalpels to cut or carve flesh. Veale (2004) used WordNet to assess the similarity of A:B to C:D as a function of the combined similarity of A to C and of B to D. In contrast, Turney (2005) used the web to pursue a more divergent course, to represent the tacit relationships of A to B and of C to D as points in a high-dimensional space. The dimensions of this space initially correspond to linking phrases on the web, before these dimensions are significantly reduced using *singular value decomposition* (SVD).

In the infamous SAT test, an analogy *A:B::C:D* has four other pairs of concepts that serve as likely

distractors (e.g. *singer:songwriter* for *hack:writer*) and the goal is to choose the most appropriate *C:D* pair for a given *A:B* pairing. Using variants of Wu and Palmer (1994) on the 374 SAT analogies of Turney (2005), Veale (2004) reports a success rate of 38–44% using only WordNet-based similarity. In contrast, Turney (2005) reports up to 55% success on the same analogies, partly because his approach aims to match implicit relations rather than explicit concepts, and in part because it uses a divergent process to gather from the web as rich a perspective as it can on these latent relationships.

2.1 Clever Comparisons Create Similarity

Each of these approaches to similarity is a *user* of information, rather than a *creator*, and each fails to capture how a creative comparison (such as a metaphor) can spur a listener to view a topic from an atypical perspective. Camac & Glucksberg (1984) provide experimental evidence for the claim that "metaphors do not use preexisting associations to achieve their effects […] people use metaphors to create new relations between concepts." They also offer a salutary reminder of an often overlooked fact: every comparison exploits information, but each is also a source of new information in its own right. Thus, "this cola is acid" reveals a different perspective on *cola* (e.g. as a *corrosive substance* or an *irritating food*) than "this acid is cola" highlights for *acid* (such as e.g., a *familiar substance*)

Veale & Keane (1994) model the role of similarity in realizing the long-term perlocutionary effect of an informative comparison. For example, to compare surgeons to butchers is to encourage one to see all surgeons as more *bloody, crude* or *careless*. The reverse comparison, of butchers to surgeons, encourages one to see butchers as more *skilled* and *precise*. Veale & Keane present a network model of memory, called *Sapper*, in which activation can spread between related concepts, thus allowing one concept to prime the properties of a neighbor. To interpret an analogy, *Sapper* lays down new activation-carrying bridges in memory between analogical counterparts, such as between *surgeon* and *butcher*, *flesh* and *meat*, or *scalpel* and *cleaver*. Comparisons thus have lasting effects on how *Sapper* sees the world, changing the pattern of activation that arises whenever it primes a concept.

Veale (2003) adopts a similarly dynamic view of similarity in WordNet, showing how an analogical comparison can result in the automatic addition of new categories and relations to WordNet itself. Veale considers the problem of finding an analogical mapping between different parts of WordNet's noun-sense hierarchy, such as between instances of *Greek god* and *Norse god*, or between the letters of different alphabets, such as of Greek and Hebrew. But no structural similarity measure for WordNet exhibits enough discernment to e.g. assign a higher similarity to *Zeus* & *Odin* (each is the supreme deity of its pantheon) than to a pairing of *Zeus* and any other *Norse god*, just as no structural measure will assign a higher similarity to *Alpha* & *Aleph* or to *Beta* & *Beth* than to any random letter pairing.

A fine-grained category hierarchy permits fine-grained similarity judgments, and though WordNet is useful, its sense hierarchies are not especially fine-grained. However, we can automatically make WordNet subtler and more discerning, by adding new fine-grained categories to unite lexical concepts whose similarity is not reflected by any existing categories. Veale (2003) shows how a property that is found in the glosses of two lexical concepts, of the same depth, can be combined with their LCS to yield a new fine-grained parent category, so e.g. "supreme" + *deity* = *Supreme-deity* (for *Odin, Zeus, Jupiter*, etc.) and "1st" + *letter* = *1st-letter* (for *Alpha, Aleph*, etc.) Selected aspects of the textual similarity of two WordNet glosses – the key to similarity in Lesk (1986) – can thus be reified into a lasting and explicitly categorical WordNet form.

3 Divergent Forms of (Re)Categorization

To tap into a richer source of concept properties than WordNet's glosses, we can use web n-grams. Consider these descriptions of a *cowboy* from the Google n-grams (Brants & Franz, 2006). The numbers to the right are Google frequency counts.

a	*lonesome*	cowboy	432
a	*mounted*	cowboy	122
a	*grizzled*	cowboy	74
a	*swaggering*	cowboy	68

To find the stable properties that can underpin a meaningful fine-grained category for *cowboy*, we must seek out the properties that are so often presupposed to be salient of all cowboys that one can use them to anchor a simile, such as "*swaggering like a cowboy*" or "*as grizzled as a cowboy*". So for each property *P* suggested by Google n-grams for a lexical concept *C*, we generate a *like*-simile for verbal behaviors such as *swaggering* and an *as-*

as-simile for adjectives such as *lonesome*. Each is then dispatched to Google as a phrasal query. We value quality over size, as these similes will later be used to find diverse viewpoints on the web via bootstrapping. We thus manually filter each web simile, to weed out any that are ill-formed, and those intended to be seen as ironic by their authors. This gives us a body of 12,000+ valid web similes.

Veale (2011, 2012, 2013) notes that web uses of the pattern "*as P as C*" are rife with irony. In contrast, web instances of "*P S such as C*" – where *S* denotes a superordinate of *C* – are rarely ironic. Hao & Veale (2010) exploit this fact to filter ironic comparisons from web similes, by re-expressing each "*as P as C*" simile as "*P * such as C*" (using a wildcard * to match any values for *S*) and looking for attested uses of this new form on the web. Since each hit will also yield a value for *S* via the wildcard *, and a fine-grained category *P-S* for *C*, we use this approach here to harvest fine-grained categories from the web from most of our similes.

Once *C* is seen to be an exemplary member of the category *P-S*, such as *cola* in *fizzy-drink*, a targeted web search is used to find other members of *P-S*, via the anchored query "*P S such as * and C*". For example, "*fizzy drinks such as * and cola*" will retrieve web texts in which * is matched to *soda* or *lemonade*. Each new member can then be used to instantiate a further query, as in "*fizzy drinks such as * and soda*", to retrieve other members of *P-S*, such as *champagne* and *root beer*. This bootstrapping process runs in successive cycles, using doubly-anchored patterns that – following Kozareva *et al.* (2008) and Veale *et al.* (2009) – explicitly mention both the category to be populated (*P-S*) and a recently acquired member of this category (*C*).

As cautioned by Kozareva *et al.*, it is reckless to bootstrap from members to categories to members again if each enfilade of queries is likely to return noisy results. A reliable filter must be applied at each stage, to ensure that any member *C* that is placed in a category *P-S* is a sensible member of the category *S*. Only by filtering in this way can we stop the rapid accumulation of noise. For instance, a WordNet-based filter can discard any categorization statement "*P S such as X and C*" where *X* does not denote a WordNet entry for which S does not denote a valid hypernym. Such a filter offers no creative latitude, however, since it forces every pairing of *C* and *P-S* to precisely obey WordNet's category hierarchy. We use instead the *near-miss* filter described in Veale *et al.* (2009), in which *X* must denote a descendant of some direct hypernym of some sense of *S*. The filter does not (and cannot) determine whether *P* is salient for *X*. It merely assumes that if *P* is salient for *C*, it is salient for *X*.

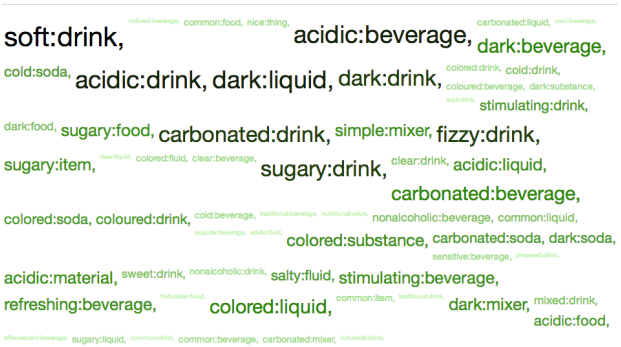

Figure 1. *Fine-grained perspectives for cola found by Thesaurus Rex on the web. See also Figures 3 and 4.*

Five successive cycles of bootstrapping are performed, using the 12,000+ web similes as a starting point. Consider *cola*: after 1 cycle, we acquire 14 new categories, such as *effervescent-beverage* and *sweet-beverage*. After 2 cycles we acquire 43 categories; after 3 cycles, 72; after 4 cycles, 93; and after 5 cycles, we acquire 102 fine-grained perspectives on *cola*, such as *stimulating-drink* and *corrosive-substance*. These alternative viewpoints, for a broad array of concepts, are gleaned from the collective intelligence of the web. Some are more discerning and informative than others – see for instance *war* & *divorce* in Figure 1 – though as de Bono (1971) notes, lateral thinking does not privilege a narrow set of "correct" viewpoints, rather it generates a broad array of interesting alternatives, none of which are ever "wrong", even if some prove more useful than others in a given context.

4 Measuring and *Creating* Similarity

Which perspectives will be most useful and informative to a WordNet-based similarity metric? Simply, a perspective *M-Cx* for a concept *Cy* can be coherently added to WordNet *iff Cx* denotes a hypernym of some sense of *Cy* in WordNet. For purposes of quantifying the similarity of two terms t_1 and t_2 – by finding the WordNet senses of these terms that exhibit the highest similarity – we can augment WordNet with the perspectives on t_1 and t_2 that are coherent with WordNet's hierarchy. So

for t_1=*cola* & t_2=*acid*, *corrosive-substance* offers a coherent new perspective on each, slotting in beneath the matching WordNet sense of *substance*.

A category system is a structured feature space. We estimate the similarity of C_1 and C_2 in WordNet as the cosine of the angle between the richest feature vectors we construct for each. The dimensions of these vectors are the atomic hypernyms (direct or indirect) of C_1 and C_2. The value of a dimension H in a feature vector is the information content (IC) of the corresponding hypernym H:

$$(1)\quad \text{IC(H)} = -\log\left(\frac{size(\text{H})}{\sum_{c \in WN} size(\text{c})}\right)$$

Here *size*(H) is the total number of lexical concepts in category *H* in WordNet, excluding any instance-level concepts, as these illustrative individuals are not evenly distributed across WordNet categories.

We also want any fine-grained perspective *M-H* to influence our similarity metric, provided it can be coherently tied into WordNet as a shared hypernym of the two lexical concepts being compared. The absolute information content of a category *M-H* that is newly added to WordNet is given by (2):

$$(2)\quad \text{IC}_{abs}(\text{M-H}) = -\log\left(\frac{size(\text{M-H})}{\sum_{m\text{-}h \in WN} size(\text{m-h})}\right)$$

where *size*(M-H) is the number of lexical concepts in WordNet for which *M-H* can be added as a new hypernym. The denominator in (2) denotes the sum total of the size of *all* fine-grained categories that can be coherently added to WordNet for *any* term.

The IC of *M-H* relative to *H* is estimated via the geometric mean of IC_{abs}(M-H) and IC(H), in (3):

$$(3)\quad \text{IC(M-H)} = \sqrt{\text{IC}_{abs}(\text{M-H}) \,.\, \text{IC(H)}}$$

For any shared dimension *H* in the feature vectors of concepts C_1 and C_2, if at least one fine-grained perspective *M-H* has been added to WordNet between *H* and C_1 and between *H* and C_2, then the value of dimension *H* for C_1 and for C_2 is given by:

$$(4)\quad weight(\text{H}) = \max(\text{IC(H)},\ \max_M \text{IC(M-H)})$$

When no shared perspective *M-H* can be added under *H*, then *weight*(H) = IC(H). A fine-grained perspective *M-H* will thus influence a similarity judgment between C_1 and C_2 only if *M-H* can be coherently added to WordNet as a hypernym of C_1 *and* C_2, and if *M-H* enriches our view of *H*. Unlike Resnick (1995), Lin (1998) and Seco *et al*. (2006), this vector-space approach does not hinge on the information content of a single LCS, so any shared hypernym *H* or perspective *M-H* can shape a similarity judgment according to its informativeness.

5 Empirical Evaluation

Many fascinating perspectives on familiar ideas are bootstrapped from the web using similes as a starting point. These perspectives drive an exploratory web-aid to lateral thinking we call *Thesaurus Rex,* while the cosine-distance metric constructed from WordNet and these many fine-grained categories is called, simply, *Rex*. When *Rex* provides a numeric estimate of similarity for two ideas, *Thesaurus Rex* provides an enhanced insight into why these ideas are similar, e.g. by explaining that *cola* & *acid* are not just substances, they are *corrosive* substances.

We evaluate *Rex* by estimating how closely its judgments correlate with those of human judges on the 30-pair word set of Miller & Charles (M&C), who aggregated the judgments of multiple human raters into mean ratings for these pairs. We evaluate three variants of *Rex* on M&C: ***Rex*-lat**, which combines WordNet with all of *Thesaurus Rex*; ***Rex*-wn**, which uses only WordNet, with nothing at all from *Thesaurus Rex*; and ***Rex*-pop**, which enriches WordNet with only *popular* perspectives from *Thesaurus Rex*. A perspective is considered popular if it is discovered *5* or more times in the bootstrapping process, using 5 different anchors. While *corrosive-substance* is a popular category for *acid*, it not so for *cola* or *juice*. Popularity thus approximates what Ortony (1979) calls *salience*.

Table 1 lists coefficients of correlation (using Pearson's ***r***) with mean human ratings for a range of WordNet-based metrics. Table 1 also includes the hybrid *WordNet+web+SVM* metric of Agirre *et al.* (2009) – who report a correlation of **.93** – and the Mutual-Information-based PMI*max* metric of Han *et al*. (2009). The latter achieves good results for 27 of the 30 M&C pairs by enriching a PMI-based metric with an automatically-generated thesaurus. While informative, this auto-generated thesaurus is not organized as an explanatory system of hierarchical categories as it is in *Thesaurus Rex*.

Similarity metric	*r*	*Similarity metric*	*r*
Wu & Palmer'94*	.74	Seco *et al.* '06*	.84
Resnick '95*	.77	Agirre *et al.* '09	**.93**
Leacock/Chod'98*	.82	Han *et al.*'09	.856
Lin '98*	.80	***Rex*-wn**	.84
Jiang/Conrath '97*	-.81	***Rex*-lat**	.89
Li *et al.* '03	.89	***Rex*-pop**	**.93**

Table 1. *Pearson product-moment correlations with mean human ratings on all 30 pairs of Miller & Charles. * as re-evaluated by Seco et al.'06 for all 30 word pairs*

***Rex*-wn** does no better than Seco *et al.* (2006) on the M&C dataset, suggesting that *Rex*'s vectors of IC-weighted hypernyms are no more discerning than a single informative LCS. However, such vectors also permit *Rex* to incorporate additional, fine-grained perspectives from *Thesaurus Rex*, allowing ***Rex*-lat** in turn to achieve a comparable correlation to that of Li *et al.* (2003) – **.89**. Yet the formulation in (2) favors unusual or idiosyncratic perspectives that are unlikely to generalize across independent judges. The mean ratings of M&C are the stuff of consensus, not individual creativity, and outside the realm of creative metaphor it often makes sense to safely align our judgments with those of others.

By limiting its use of *Thesaurus Rex* to the perspectives that other judges are most likely to use, ***Rex*-pop** obtains a correlation of .**93** with mean human ratings on all 30 M&C pairs. This result is comparable to that reported by Agirre *et al*. (2009), who use SVM-based supervised learning to combine the judgments of two metrics, one based on WordNet and another on the analysis of web contexts of both input terms. However, *Rex* has the greater capacity for insight, since it augments the structured category system of WordNet with structured categories of its own. At each level of the WordNet hierarchy, *Rex* finds the fine-grained category that can best inform its judgments. Because *Rex* makes highly selective use of the diverse products of lateral thinking, this selectivity also produces concise explanations for its judgments.

5.1 Generative Uses of Similarity

A similarity metric offers a numerical measure of how closely one idea can cluster with another. It can also indicate how well one object may serve as a substitute for another, as when a *letter opener* is used as a *knife*, or *tofu* is used instead of *meat*. This need for substitution can be grist for creativity, yet most similarity metrics can only assess a suggested substitution, rather than suggest one themselves. If they are to actively shape a creative decision, our similarity metrics must be made more generative.

A similarity metric can learn to be generative, by observing how people typically cluster words and ideas that are made similar by their contexts of use. The Google 3-grams contain many instances of the clustering pattern "*X+s and Y+s*", as in "cowboys and pirates" or "doctors and lawyers", and so a comprehensive trawl yields many insights into the pairings of ideas that we implicitly see as comparable. We harvest all such Google 3-grams, to build a symmetric *comparability graph* in which any two comparable terms are adjacent nodes. For any node, we can generate a diverse set of comparable ideas just by reading off its adjacent nodes. *Thesaurus Rex* can be used to find an embracing category for many such pairs of nodes, while *Rex* estimates the similarity of any two adjacent nodes. A comparability graph of 28,000 nodes is produced from the Google 3-grams, with a sparse adjacency matrix of just 1,264,827 (0.16%) non-zero entries.

Is this dense enough for a task requiring generative similarity? Almuhareb & Poesio (2004) describe one such task: they sample 214 words from across 13 WordNet categories, and ask if these 214 words can be partitioned into 13 clusters that mirror the WordNet categories from which they were drawn. They then collect tens of thousands of web contexts for these 214 words, to extract a feature representation of each. We instead use *Rex* to generate, as features, a diverse set of comparable terms for each word. (We also assume that each word is a feature of itself). The *Rex* comparability graph suggests a pool of 8,300 features for all 214 words. The clustering toolkit CLUTO is used to partition the original 214 words into 13 clusters guided only by these comparability features. The resulting 13 clusters have an average purity of **93.4%** relative to WordNet, suggesting that categorization tasks which require implicit comparability judgments are well served by a generative approach to similarity.

5.2 Learning From Similarity Judgments

Rex augments the narrow worldview of WordNet with the more diverse viewpoints it gleans from the web, not by viewing them as separate knowledge sources, but by actually updating WordNet itself. The relative performance of ***Rex*-pop** > ***Rex*-lat** >

***Rex*-wn** on the M&C dataset shows that selective use of a divergent perspective permits WordNet to better serve its popular role as a judge of similarity. It is worth asking then whether these passing additions to WordNet should not be made permanent.

Rex estimates a similarity score for each of the 1,264,827 pairings of comparable terms it finds in the Google 3-grams. These scores are then cached to support generative similarity, and to permit fast lookup of scores for common comparisons. This lookup table is a lightweight means of using *Rex* in a range of creative substitution or generation tasks. Though the table is sparse, §5.1 shows that it implicitly captures key nuances of category structure. The 39,826 unique fine-grained categories added by ***Rex*-pop** (versus the 44,238 categories added by ***Rex*-lat**) in the course of its 1,264,827 comparisons thus suggest credible enhancements to WordNet. Figure 2 graphs the distribution of new categories and their sizes when ***Rex*-pop** is used on this scale.

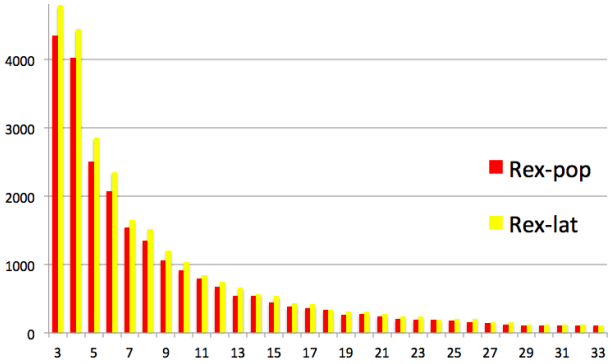

Figure 2. *The no. of new categories of a given size added to WordNet when **Rex-pop/lat** are used at web scale.*

The *Goldilocks categories* are those that are not so small as to lack generality, and not so large as to lack information content. For example, ***Rex*-pop** suggests the addition of 15,125 new fine-grained categories to WordNet with membership sizes ranging from 5 to 25. This is a large but manageable number of categories that should be further considered for future addition to WordNet, or indeed to any similarly curated knowledge resource.

6 Summary and Conclusions

de Bono (1970) argues that the best solutions arise from using lateral and vertical thinking *in unison*. Lateral thinking is divergent and generative, while vertical thinking is convergent and analytical. The former can thus be used to create a pool of interesting candidates for the latter to selectively consider. *Thesaurus Rex* uses the web to generate a rich pool of alternate perspectives on familiar ideas, and *Rex* selects from this pool to perform vertical reasoning with WordNet to yield precise similarity judgments. *Rex* also uses the most informative perspective to concisely *explain* each comparison, or – when used in generative mode – to *suggest* a creative comparison. For instance, to highlight the potential toxicity of *coffee*, *Thesaurus Rex* suggests comparisons with *alcohol*, *tobacco* or *pesticide*, as all have been categorized as toxic substances on the web. A web app based on *Thesaurus Rex*, to support this kind of lateral thinking, is accessible online at this URL:

http://boundinanutshell.com/therex2

Screenshots from the *Thesaurus Rex* application are provided in Figures 3 and 4 overleaf. Because *Thesaurus Rex* targets the acquisition of fine-grained perspectives, ranging from the offbeat to the obvious, it acquires an order-of-magnitude more categories from the web than can be found in WordNet itself. *Rex* dips selectively into this wealth of perspectives (and ***Rex*-pop** is more selective still), though many of *Rex*'s needs can be anticipated by looking to how ideas are implicitly grouped into *ad-hoc categories* (Barsalou, 1983) in constructions such as "*X+s and Y+s*". Using the Google n-grams as a source of tacit grouping constructions, we have created a comprehensive lookup table that provides *Rex* similarity scores for the most common (if often implicit) comparisons.

Comparability is not the same as similarity, and a non-zero similarity score does not mean that two concepts would ever be considered comparable by a human. This poses a problem for the generation of sensible comparisons. However, *Rex*'s lookup table captures the implicit pragmatics of comparability, making *Rex* usable in generative tasks where a metric must both suggest and evaluate comparisons. Human similarity mechanisms are evaluative *and* generative, convergent *and* divergent. Our computational mechanisms should be no less so.

7 Acknowledgements

This research was partly supported by the WCU (World Class University) program under the National Research Foundation of Korea (Ministry of Education, Science and Technology of Korea, Project no. R31-30007) and partly funded by Science Foundation Ireland via the *Centre for Next Generation Localization* (CNGL).

Figure 3. A screenshot from the web application *Thesaurus Rex,* showing the fine-grained categories found by *Thesaurus Rex* for the lexical concept ***creativity*** on the web.

Figure 4. A screenshot from the web application *Thesaurus Rex,* showing the shared overlapping categories found by *Thesaurus Rex* for the lexical concepts ***divorce*** and ***war***.

References

Aristotle (translator: James Hutton). 1982. *Aristotle's Poetics*. New York: Norton.

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca and Aitor Soroa. 2009. Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. *In Proceedings of NAACL '09, The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 19—27.

Abdulrahman Almuhareb and Massimo Poesio. 2004. Attribute-Based and Value-Based Clustering: An Evaluation. In *Proceedings of the Conference on Empirical Methods in NLP,* Barcelona. pp. 158-165.

Lawrence W. Barsalou. 1983. Ad hoc categories. Memory and Cognition, 11:211–227.

Thorsten Brants and Alex Franz. 2006. *Web 1T 5-gram Ver. 1*. Philadelphia: Linguistic Data Consortium.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics,* 32(1):13-47.

Mary K. Camac, and Sam Glucksberg. 1984. Metaphors do not use associations between concepts, they are used to create them. *Journal of Psycholinguistic Research*, **13**, 443-455.

de Bono, Edward. 1970. *Lateral thinking: creativity step by step*. New York: Harper & Row.

de Bono, Edward. 1971. *Lateral thinking for management: a handbook for creativity*. New York: McGraw Hill.

Christiane Fellbaum (ed.). 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

J. Paul Guilford. 1967. *The Nature of Human Intelligence*. New York: McGraw Hill.

Lushan Han, Tim Finin, Paul McNamee, Anupam Joshi and Yelena Yesha. 2012. Improving Word Similarity by Augmenting PMI with Estimates of Word Polysemy. *IEEE Transactions on Data and Knowledge Engineering* (13 Feb. 2012).

Yanfen Hao and Tony Veale. 2010. An Ironic Fist in a Velvet Glove: Creative Mis-Representation in the Construction of Ironic Similes. *Minds and Machines* **20**(4), pp. 635–650.

Jay Y. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, pp. 19-33.

Zornitsa Kozareva, Eileen Riloff and Eduard Hovy. 2008. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. *In Proc. of the 46th Annual Meeting of the ACL,* pp 1048-1056.

Claudia Leacock and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database,* 265–283.

Yuhua Li, Zuhair A. Bandar and David McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882.

Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th ICML, the International Conference on Machine Learning*, Morgan Kaufmann, San Francisco CA, pp. 296– 304.

Michael Lesk. 1986 Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SigDoc, ACM*, 24–26.

George A. Miller and Walter. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* **6**(1):1-28.

Andrew Ortony. 1979. Beyond literal similarity. *Psychological Review, 86,* pp. 161-180.

Ted Pederson, Siddarth Patwardhan and Jason Michelizzi. 2004. WordNet::Similarity: measuring the relatedness of concepts. In *Proceedings of HLT-NAACL'04 (Demonstration Papers) the 2004 annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 38-41.*

Philip Resnick. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of IJCAI'95, the 14th International Joint Conference on Artificial Intelligence.*

Nuno Seco, Tony Veale and Jer Hayes, 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *Proceedings of ECAI'04*, *the European Conference on Artificial Intelligence.*

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proceedings of AAAI-06, the 2006 Conference of the Association for the Advancement of AI,* pp. 1419–1424.

Peter Turney. 2005. Measuring semantic similarity by latent relational analysis. *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, 1136-1141.

Tony Veale and Mark T. Keane. 1994. Belief Modeling, Intentionality and Perlocution in Metaphor Comprehension. In Proceedings of the 16th Annual Meeting of the Cognitive Science Society, Atlanta, Georgia. Hillsdale, NJ: Lawrence Erlbaum.

Tony Veale. 2003. The analogical thesaurus: An emerging application at the juncture of lexical metaphor and information retrieval. In *Proceedings of IAAI'03, the 15th International Conference on Innovative Applications of Artificial Intelligence*, Mexico.

Tony Veale. 2004. WordNet sits the SAT: A knowledge-based approach to lexical analogy. *Proceedings of ECAI'04, the European Conference on Artificial Intelligence*, 606-612.

Tony Veale and Yanfen Hao. 2007. Comprehending and Generating Apt Metaphors: A Web-driven, Case-based Approach to Figurative Language. In proceedings of AAAI 2007, the 22nd AAAI Conference on Artificial Intelligence. Vancouver, Canada.

Tony Veale, Guofu Li and Yanfen Hao. 2009. Growing Finely-Discriminating Taxonomies from Seeds of Varying Quality and Size. In *Proc. of EACL'09, the 12th Conference of the European Chapter of the Association for Computational Linguistics* pp. 835-842.

Tony Veale. 2011. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. In Proceedings of ACL'2011, the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.

Tony Veale. 2012. Exploding the Creativity Myth: The computational foundations of linguistic creativity. *London: Bloomsbury Academic.*

Tony Veale. 2013. Humorous Similes. *Humor: The International Journal of Humor Research, **21**(1):3-22.*

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of ACL'94, 32nd annual meeting of the Association for Computational Linguistics, Las Cruces, New Mexico,*. pp. 133-138.