

Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity

Tony Veale

School of Computer Science and Informatics,
University College Dublin,
Belfield, Dublin D4, Ireland.

Tony.Veale@UCD.ie

Abstract

Information retrieval (IR) and figurative language processing (FLP) could scarcely be more different in their treatment of language and meaning. IR views language as an open-ended set of mostly stable signs with which texts can be indexed and retrieved, focusing more on a text's potential relevance than its potential meaning. In contrast, FLP views language as a system of *unstable* signs that can be used to talk about the world in creative new ways. There is another key difference: IR is practical, scalable and robust, and in daily use by millions of casual users. FLP is neither scalable nor robust, and not yet practical enough to migrate beyond the lab. This paper thus presents a mutually beneficial hybrid of IR and FLP, one that enriches IR with new operators to enable the non-literal retrieval of creative expressions, and which also transplants FLP into a robust, scalable framework in which practical applications of linguistic creativity can be implemented.

1 Introduction

Words should not always be taken at face value. Figurative devices like metaphor can communicate far richer meanings than are evident from a superficial – and perhaps *literally* nonsensical – reading. Figurative Language Processing (FLP) thus uses a variety of special mechanisms and representations,

to assign non-literal meanings not just to metaphors, but to similes, analogies, epithets, puns and other creative uses of language (see Martin, 1990; Fass, 1991; Way, 1991; Indurkha, 1992; Fass, 1997; Barnden, 2006; Veale and Butnariu, 2010).

Computationalists have explored heterodox solutions to the procedural and representational challenges of metaphor, and FLP more generally, ranging from flexible representations (e.g. the *preference semantics* of Wilks (1978) and the *collative semantics* of Fass (1991, 1997)) to processes of cross-domain structure alignment (e.g. *structure mapping theory*; see Gentner (1983) and Falkenhainer *et al.* 1989) and even structural inversion (Veale, 2006). Though thematically related, each approach to FLP is broadly distinct, giving computational form to different cognitive demands of creative language: thus, some focus on inter-domain mappings (e.g. Gentner, 1983) while others focus more on intra-domain inference (e.g. Barnden, 2006). However, while computationally interesting, none has yet achieved the scalability or robustness needed to make a significant practical impact outside the laboratory. Moreover, such systems tend to be developed in isolation, and are rarely designed to cohere as part of a larger framework of creative reasoning (e.g. Boden, 1994).

In contrast, Information Retrieval (IR) is both scalable and robust, and its results translate easily from the laboratory into practical applications (e.g. see Salton, 1968; Van Rijsbergen, 1979). Whereas FLP derives its utility *and* its fragility from its attempts to identify deeper meanings beneath the surface, the widespread applicability of IR stems directly from its superficial treatment of language

and meaning. IR does not distinguish between creative and conventional uses of language, or between literal and non-literal meanings. IR is also remarkably modular: its components are designed to work together interchangeably, from stemmers and indexers to heuristics for query expansion and document ranking. Yet, because IR treats all language as literal language, it relies on literal matching between queries and the texts that they retrieve. Documents are retrieved precisely because they contain stretches of text that literally resemble the query. This works well in the main, but it means that IR falls flat when the goal of retrieval is not to identify relevant documents but to retrieve new and creative ways of expressing a given idea. To retrieve creative language, and to be potentially surprised or inspired by the results, one needs to facilitate a non-literal relationship between queries and the texts that they match.

The complementarity of FLP and IR suggests a productive hybrid of both paradigms. If the most robust elements of FLP are used to provide new non-literal query operators for IR, then IR can be used to retrieve potentially new and creative ways of speaking about a topic from a large text collection. In return, IR can provide a stable, robust and extensible platform on which to use these operators to build FLP systems that exhibit linguistic creativity. In the next section we consider the related work on which the current realization of these ideas is founded, before presenting a specific trio of new semantic query operators in section 3. We describe three simple but practical applications of this creative IR paradigm in section 4. Empirical support for the FLP intuitions that underpin our new operators is provided in section 5. The paper concludes with some closing observations about future goals and developments in section 6.

2 Related Work and Ideas

IR works on the premise that a user can turn an information need into an effective query by anticipating the language that is used to talk about a given topic in a target collection. If the collection uses creative language in speaking about a topic, then a query must also contain the seeds of this creative language. Veale (2004) introduces the idea of *creative information retrieval* to explore how an IR system can itself provide a degree of creative anticipation, acting as a mediator between the lit-

eral specification of a meaning and the retrieval of creative articulations of this meaning. This anticipation ranges from simple re-articulation (e.g. a text may implicitly evoke “*Qur’an*” even if it only contains “*Muslim bible*”) to playful allusions and epithets (e.g. the CEO of a rubber company may be punningly described as a “*rubber baron*”). A creative IR system may even anticipate out-of-dictionary words, like *chocoholic* and *sexoholic*.

Conventional IR systems use a range of query expansion techniques to automatically bolster a user’s query with additional keywords or weights, to permit the retrieval of relevant texts it might not otherwise match (e.g. Vernimb, 1977; Voorhees, 1994). Techniques vary, from the use of stemmers and morphological analysis to the use of thesauri (such as WordNet; see Fellbaum, 1998; Voorhees, 1998) to pad a query with synonyms, to the use of statistical analysis to identify more appropriate context-sensitive associations and near-synonyms (e.g. Xu and Croft, 1996). While some techniques may suggest conventional metaphors that have become lexicalized in a language, they are unlikely to identify relatively novel expressions. Crucially, expansion improves recall at the expense of overall precision, making automatic techniques even more dangerous when the goal is to retrieve results that are creative *and* relevant. Creative IR must balance a need for fine user control with the statistical breadth and convenience of automatic expansion.

Fortunately, statistical corpus analysis is an obvious area of overlap for IR and FLP. Distributional analyses of large corpora have been shown to produce nuanced models of lexical similarity (e.g. Weeds and Weir, 2005) as well as context-sensitive thesauri for a given domain (Lin, 1998). Hearst (1992) shows how a pattern like “*Xs and other Ys*” can be used to construct more fluid, context-specific taxonomies than those provided by WordNet (e.g. “*athletes and other celebrities*” suggests a context in which athletes are viewed as stars). Mason (2004) shows how statistical analysis can automatically detect and extract conventional metaphors from corpora, though creative metaphors still remain a tantalizing challenge. Hanks (2005) shows how the “*Xs like A, B and C*” construction allows us to derive flexible ad-hoc categories from corpora, while Hanks (2006) argues for a gradable conception of metaphoricity based on word-sense distributions in corpora.

Veale and Hao (2007) exploit the simile frame “*as X as Y*” to harvest a great many common similes and their underlying stereotypes from the web (e.g. “*as hot as an oven*”), while Veale and Hao (2010) show that the pattern “*about as X as Y*” retrieves an equally large collection of creative (if mostly ironic) comparisons. These authors demonstrate that a large vocabulary of stereotypical ideas (over 4000 nouns) and their salient properties (over 2000 adjectives) can be harvested from the web.

We now build on these results to develop a set of new semantic operators, that use corpus-derived knowledge to support finely controlled non-literal matching and automatic query expansion.

3 Creative Text Retrieval

In language, creativity is always a matter of construal. While conventional IR queries articulate a need for information, creative IR queries articulate a need for expressions to convey the same meaning in a fresh or unusual way. A query and a matching phrase can be figuratively construed to have the same meaning if there is a non-literal mapping between the elements of the query and the elements of the phrase. In creative IR, this non-literal mapping is facilitated by the query’s explicit use of *semantic wildcards* (e.g. see Mihalcea, 2002).

The wildcard * is a boon for power-users of the Google search engine, precisely because it allows users to focus on the retrieval of matching phrases rather than relevant documents. For instance, * can be used to find alternate ways of instantiating a culturally-established linguistic pattern, or “snowclone”: thus, the Google queries “*In * no one can hear you scream*” (from *Alien*), “*Reader, I * him*” (from *Jane Eyre*) and “*This is your brain on **” (from a famous TV advert) find new ways in which old patterns have been instantiated for humorous effect on the Web. On a larger scale, Veale and Hao (2007) used the * wildcard to harvest web similes, but reported that harvesting cultural data with wildcards is not a straightforward process. Google and other engines are designed to maximize document relevance and to rank results accordingly. They are not designed to maximize the diversity of results, or to find the largest set of wildcard bindings. Nor are they designed to find the most commonplace bindings for wildcards.

Following Guilford’s (1950) pioneering work, diversity is widely considered a key component in

the psychology of creativity. By focusing on the phrase level rather than the document level, and by returning phrase sets rather than document sets, creative IR maximizes diversity by finding as many bindings for its wildcards as a text collection will support. But we need more flexible and precise wildcards than *. We now consider three varieties of semantic wildcards that build on insights from corpus-linguistic approaches to FLP.

3.1 The Neighborhood Wildcard ?X

Semantic query expansion replaces a query term **X** with a set {**X**, **X**₁, **X**₂, ..., **X**_n} where each **X**_i is related to **X** by a prescribed lexico-semantic relationship, such as synonymy, hyponymy or meronymy. A generic, lightweight resource like WordNet can provide these relations, or a richer ontology can be used if one is available (e.g. see Navigli and Velardi, 2003). Intuitively, each query term suggests other terms from its semantic neighborhood, yet there are practical limits to this intuition. **X**_i may not be an obvious or natural substitute for **X**. A neighborhood can be drawn too small, impacting recall, or too large, impacting precision.

Corpus analysis suggests an approach that is both semantic *and* pragmatic. As noted in Hanks (2005), languages provide constructions for building ad-hoc sets of items that can be considered comparable in a given context. For instance, a co-ordination of bare plurals suggests that two ideas are related at a generic level, as in “*priests and imams*” or “*mosques and synagogues*”. More generally, consider the pattern “*X and Y*”, where **X** and **Y** are proper-names (e.g., “*Zeus and Hera*”), or **X** and **Y** are inflected nouns or verbs with the same inflection (e.g., the plurals “*cats and dogs*” or the verb forms “*kicking and screaming*”). Millions of matches for this pattern can be found in the Google 3-grams (Brants and Franz, 2006), allowing us to build a map of comparable terms by linking the root-forms of **X** and **Y** with a similarity score obtained via a WordNet-based measure (e.g. see Budanitsky and Hirst (2006) for a good selection).

The pragmatic neighborhood of a term **X** can be defined as {**X**, **X**₁, **X**₂, ..., **X**_n}, so that for each **X**_i, the Google 3-grams contain “*X+inf and X_i+inf*” or “*X+inf and X_i+inf*”. The boundaries of neighborhoods are thus set by usage patterns: if **?X** denotes the neighborhood of **X**, then **?artist**

matches not just *artist, composer* and *poet*, but *studio, portfolio* and *gallery*, and many other terms that are semantically dissimilar but pragmatically linked to *artist*. Since each $X_i \in ?X$ is ranked by similarity to X , query matches can also be ranked by similarity.

When X is an adjective, then $?X$ matches any element of $\{X, X_1, X_2, \dots, X_n\}$, where each X_i pragmatically reinforces X , and X pragmatically reinforces each X_i . To ensure X and X_i really are mutually reinforcing adjectives, we use the double-ground simile pattern “*as X and X_i as*” to harvest $\{X_1, \dots, X_n\}$ for each X . Moreover, to maximize recall, we use the Google API (rather than Google ngrams) to harvest suitable bindings for X and X_i from the web. For example, $@witty = \{charming, clever, intelligent, entertaining, \dots, edgy, fun\}$.

3.2 The Cultural Stereotype Wildcard @X

Dickens claims in *A Christmas Carol* that “the wisdom of a people is in the simile”. Similes exploit familiar stereotypes to describe a less familiar concept, so one can learn a great deal about a culture and its language from the similes that have the most currency (Taylor, 1954). The wildcard $@X$ builds on the results of Veale and Hao (2007) to allow creative IR queries to retrieve matches on the basis of cultural expectations. This foundation provides a large set of adjectival features (over 2000) for a larger set of nouns (over 4000) denoting stereotypes for which these features are salient.

If N is a noun, then $@N$ matches any element of the set $\{A_1, A_2, \dots, A_n\}$, where each A_i is an adjective denoting a stereotypical property of N . For example, $@diamond$ matches any element of $\{transparent, immutable, beautiful, tough, expensive, valuable, shiny, bright, lasting, desirable, strong, \dots, hard\}$. If A is an adjective, then $@A$ matches any element of the set $\{N_1, N_2, \dots, N_n\}$, where each N_i is a noun denoting a stereotype for which A is a culturally established property. For example, $@tall$ matches any element of $\{giraffe, skyscraper, tree, redwood, tower, sunflower, lighthouse, beanstalk, rocket, \dots, supermodel\}$.

Stereotypes crystallize in a language as clichés, so one can argue that stereotypes and clichés are little or no use to a creative IR system. Yet, as demonstrated in Fishlov (1992), creative language

is replete with stereotypes, not in their clichéd guises, but in novel and often incongruous combinations. The creative value of a stereotype lies in how it is used, as we’ll show later in section 4.

3.3 The Ad-Hoc Category Wildcard ^X

Barsalou (1983) introduced the notion of an *ad-hoc category*, a cross-cutting collection of often disparate elements that cohere in the context of a specific task or goal. The ad-hoc nature of these categories is reflected in the difficulty we have in naming them concisely: the cumbersome “things to take on a camping trip” is Barsalou’s most cited example. But ad-hoc categories do not replace natural kinds; rather, they supplement an existing system of more-or-less rigid categories, such as the categories found in WordNet.

The semantic wildcard C matches C and any element of $\{C_1, C_2, \dots, C_n\}$, where each C_i is a member of the category named by C . C can denote a fixed category in a resource like WordNet or even Wikipedia; thus, fruit matches any member of $\{apple, orange, pear, \dots, lemon\}$ and animal any member of $\{dog, cat, mouse, \dots, deer, fox\}$.

Ad-hoc categories arise in creative IR when the results of a query – or more specifically, the bindings for a query wildcard – are funneled into a new user-defined category. For instance, the query “ $^fruit\ juice$ ” matches any phrase in a text collection that denotes a named fruit juice, from “*lemon juice*” to “*pawpaw juice*”. A user can now funnel the bindings for fruit in this query into an ad-hoc category **juicefruit**, to gather together those fruits that are used for their juice. Elements of juicefruit are ranked by the corpus frequencies discovered by the original query; low-frequency **juicefruit** members in the Google ngrams include *coffee, raisin, almond, carob* and *soybean*. Ad-hoc categories allow users of IR to remake a category system in their own image, and create a new vocabulary of categories to serve their own goals and interests, as when “ $^food\ pizza$ ” is used to suggest disparate members for the ad-hoc category **pizzatopping**.

The more subtle a query, the more disparate the elements it can funnel into an ad-hoc category. We now consider how basic semantic wildcards can be combined to generate even more diverse results.

3.4 Compound Operators

Each wildcard maps a query term onto a set of ex-

pansion terms. The compositional semantics of a wildcard combination can thus be understood in set-theoretic terms. The most obvious and useful combinations of ?, @ and ^ are described below:

?? Neighbor-of-a-neighbor: if ?X matches any element of {X, X₁, X₂, ..., X_n} then ??X matches any of ?X ∪ ?X₁ ∪ ... ∪ ?X_n, where the ranking of X_{ij} in ??X is a function of the ranking of X_i in ?X and the ranking of X_{ij} in ?X_i. Thus, ??**artist** matches far more terms than ?**artist**, yielding more diversity, more noise, and more creative potential.

@@ Stereotype-of-a-stereotype: if @X matches any element of {X₁, X₂, ..., X_n} then @@X matches any of @X₁ ∪ @X₂ ∪ ... ∪ @X_n. For instance, @@**diamond** matches any stereotype that shares a salient property with *diamond*, and @@**sharp** matches any salient property of any noun for which *sharp* is a stereotypical property.

?@ Neighborhood-of-a-stereotype: if @X matches any element of {X₁, X₂, ..., X_n} then ?@X matches any of ?X₁ ∪ ?X₂ ∪ ... ∪ ?X_n. Thus, ?@**cunning** matches any term in the pragmatic neighborhood of a stereotype for *cunning*, while ?@**knife** matches any property that mutually reinforces any stereotypical property of *knife*.

@? Stereotypes-in-a-neighborhood: if ?X matches any of {X, X₁, X₂, ..., X_n} then @?X matches any of @X ∪ @X₁ ∪ ... ∪ @X_n. Thus, @?**corpse** matches any salient property of any stereotype in the neighborhood of *corpse*, while @?**fast** matches any stereotype noun with a salient property that is similar to, and reinforced by, *fast*.

?^ Neighborhood-of-a-category: if ^C matches any of {C, C₁, C₂, ..., C_n} then ?^C matches any of ?C ∪ ?C₁ ∪ ... ∪ ?C_n.

^? Categories-in-a-neighborhood: if ?X matches any of {X, X₁, X₂, ..., X_n} then ^?X matches any of ^X ∪ ^X₁ ∪ ... ∪ ^X_n.

@^ Stereotypes-in-a-category: if ^C matches any of {C, C₁, C₂, ..., C_n} then @^C matches any of @C ∪ @C₁ ∪ ... ∪ @C_n.

^@ Members-of-a-stereotype-category: if @X matches any element of {X₁, X₂, ..., X_n} then ^@X matches any of ^X₁ ∪ ^X₂ ∪ ... ∪ ^X_n.

So ^@**strong** matches any member of a category (such as *warrior*) that is stereotypically *strong*.

4 Applications of Creative Retrieval

The Google ngrams comprise a vast array of extracts from English web texts, of 1 to 5 words in length (Brants and Franz, 2006). Many extracts are well-formed phrases that give lexical form to many different ideas. But an even greater number of ngrams are not linguistically well-formed. The Google ngrams can be seen as a *lexicalized idea space*, embedded within a larger sea of noise. Creative IR can be used to explore this idea space.

Each creative query is a jumping off point in a space of lexicalized ideas that is implied by a large corpus, with each successive match leading the user deeper into the space. By turning matches into queries, a user can perform a creative exploration of the space of phrases and ideas (see Boden, 1994) while purposefully sidestepping the noise of the Google ngrams. Consider the pleonastic query “*Catholic ?pope*”. Retrieved phrases include, in descending order of lexical similarity, “*Catholic president*”, “*Catholic politician*”, “*Catholic king*”, “*Catholic emperor*” and “*Catholic patriarch*”. Suppose a user selects “*Catholic king*”: the new query “*Catholic ?king*” now retrieves “*Catholic queen*”, “*Catholic court*”, “*Catholic knight*”, “*Catholic kingdom*” and “*Catholic throne*”. The subsequent query “*Catholic ?kingdom*” in turn retrieves “*Catholic dynasty*” and “*Catholic army*”, among others. In this way, creative IR allows a user to explore the text-supported ramifications of a metaphor like *Popes are Kings* (e.g., if popes are kings, they too might have queens, command armies, found dynasties, or sit on thrones).

Creative IR gives users the tools to conduct their own explorations of language. The more wildcards a query contains, the more degrees of freedom it offers to the explorer. Thus, the query “*?scientist 's ?laboratory*” uncovers a plethora of analogies for the relationship between scientists and their labs: matches in the Google 3-grams include “*technician's workshop*”, “*artist's studio*”, “*chef's kitchen*” and “*gardener's greenhouse*”.

4.1 Metaphors with *Aristotle*

For a term X , the wildcard $?X$ suggests those other terms that writers have considered to be comparable to X , while $??X$ extrapolates beyond the corpus evidence to suggest an even larger space of potential comparisons. A meaningful metaphor can be constructed for X by framing X with any stereotype to which it is pragmatically comparable, that is, any stereotype in $?X$. Collectively, these stereotypes can impart the properties $@?X$ to X .

Suppose one wants to metaphorically ascribe the property P to X . The set $@P$ contains those stereotypes for which P is culturally salient. Thus, close metaphors for X (what MacCormac (1985) dubs *epiphors*) in the context of P are suggested by $?X \cap @P$. More distant metaphors (MacCormac dubs these *diaphors*) are suggested by $??X \cap @P$. For instance, to describe a *scholar* as *wise*, one can use *poet*, *yogi*, *philosopher* or *rabbi* as comparisons. Yet even a simple metaphor will impart other features to a topic. If P_S denotes the ad-hoc set of additional properties that may be inferred for X when a stereotype S is used to convey property P , then $^P_S = ?P \cap @@P$. The query “ $^P_S X$ ” now finds corpus-attested elements of P_S that can meaningfully be used to modify X .

These IR formulations are used by *Aristotle*, an online metaphor generator, to generate targeted metaphors that highlight a property P in a topic X . *Aristotle* uses the Google ngrams to supply values for $?X$, $??X$, $?P$ and P_S . The system can be accessed at: www.educatedinsolence.com/aristotle

4.2 Expressing Attitude with *Idiom Savant*

Our retrieval goals in IR are often affective in nature: we want to find a way of speaking about a topic that expresses a particular sentiment and carries a certain tone. However, affective categories are amongst the most cross-cutting structures in language. Words for disparate ideas are grouped according to the sentiments in which they are generally held. We respect *judges* but dislike *critics*; we respect *heroes* but dislike *killers*; we respect *sharpshooters* but dislike *snipers*; and respect *rebels* but dislike *insurgents*. It seems therefore that the particulars of sentiment are best captured by a set of culture-specific ad-hoc categories.

We thus construct two ad-hoc categories,

posword and negword , to hold the most obviously positive or negative words in Whissell’s (1989) *Dictionary of Affect*. We then grow these categories to include additional reinforcing elements from their pragmatic neighborhoods, $?^posword$ and $?^negword$. As these categories grow, so too do their neighborhoods, allowing a simple semi-automated bootstrapping process to significantly grow the categories over several iterations. We construct two phrasal equivalents of these categories, posphrase and negphrase , using the queries “ $^posword - ^pastpart$ ” (e.g., matching “*high-minded*” and “*sharp-eyed*”) and “ $^negword - ^pastpart$ ” (e.g., matching “*flat-footed*” and “*dead-eyed*”) to mine affective phrases from the Google 3-grams. The resulting ad-hoc categories (of ~600 elements each) are manually edited to fix any obvious mis-categorizations.

Idiom Savant is a web application that uses posphrase and negphrase to suggest flattering and insulting epithets for a given topic. The query “ $^posphrase ?X$ ” retrieves phrases for a topic X that put a positive spin on a related topic to which X is sometimes compared, while “ $^negphrase ?X$ ” conversely imparts a negative spin. Thus, for *politician*, the Google 4-grams provide the flattering epithets “*much-needed leader*”, “*awe-inspiring leader*”, “*hands-on boss*” and “*far-sighted statesman*”, as well as insults like “*power-mad leader*”, “*back-stabbing boss*”, “*ice-cold technocrat*” and “*self-promoting hack*”. Riskier diaphors can be retrieved via “ $^posphrase ??X$ ” and “ $^negphrase ??X$ ”. *Idiom Savant* is accessible online at: www.educatedinsolence.com/idiom-savant/

4.3 Poetic Similes with *The Jigsaw Bard*

The well-formed phrases of a large corpus can be viewed as the linguistic equivalent of *objets trouvés* in art: readymade or “found” objects that might take on fresh meanings in a creative context. The phrase “*robot fish*”, for instance, denotes a more-or-less literal object in the context of autonomous robotic submersibles, but can also be used to convey a figurative meaning as part of a creative comparison (e.g., “*he was as cold as a robot fish*”).

Fishlov (1992) argues that poetic comparisons are most resonant when they combine mutually-reinforcing (if distant) ideas, to create memorable images and evoke nuanced feelings. Building on Fishlov’s argument, creative IR can be used to turn

the readymade phrases of the Google ngrams into vehicles for creative comparison. For a topic X and a property P , simple similes of the form “ X is as P as S ” are easily generated, where $S \in @P \cap ??X$.

Fishlov would dub these *non-poetic similes* (NPS). However, the query “ $?P @P$ ” will retrieve corpus-attested elaborations of stereotypes in $@P$ to suggest similes of the form “ X is as P as $P_1 S$ ”, where $P_1 \in ?P$. These similes exhibit elements of what Fishlov dubs *poetic similes* (PS). Why say “as cold as a fish” when you can say “as cold as a wet fish”, “a dead haddock”, “a wet January”, “a frozen corpse”, or “a heartless robot”? Complex queries can retrieve more creative combinations, so “ $@P @P$ ” (e.g. “robot fish” or “snow storm” for cold), “ $?P @P @P$ ” (e.g. “creamy chocolate mousse” for rich) and “ $@P - ^{pastpart} @P$ ” (e.g. “snow-covered graveyard” and “bullet-riddled corpse” for cold) each retrieve ngrams that blend two different but overlapping stereotypes.

Blended properties also make for nuanced similes of the form “as P and $?P$ as S ”, where $S \in @P \cap @?P$. While one can be “as rich as a fat king”, something can be “as rich and enticing as a chocolate truffle”, “a chocolate brownie”, “a chocolate fruitcake”, and even “a chocolate king”.

The *Jigsaw Bard* is a web application that harnesses the readymades of the Google ngrams to formulate novel similes from existing phrases. By mapping blended properties to ngram phrases that combine multiple stereotypes, the *Bard* expands its generative scope considerably, allowing this application to generate hundreds of thousands of evocative comparisons. The *Bard* can be accessed online at: www.educatedinsolence.com/jigsaw/

5 Empirical Evaluation

Though $^$ is the most overtly categorical of our wildcards, all three wildcards – $?$, $@$ and $^$ – are categorical in nature. Each has a semantic or pragmatic membership function that maps a term onto an expansion set of related members. The membership functions for specific uses of $^$ are created in an ad-hoc fashion by the users that exploit it; in contrast, the membership functions for uses of $@$ and $?$ are derived automatically, via pattern-matching and corpus analysis. Nonetheless, ad-hoc categories in creative IR are often populated with the bindings produced by uses of $@$ and

$?$ and combinations thereof. In a sense, $?X$ and $@X$ and their variations are themselves ad-hoc categories. But how well do they serve as categories? Are they large, but noisy? Or too small, with limited coverage? We can evaluate the effectiveness of $?$ and $@$, and indirectly that of $^$ too, by comparing the use of $?$ and $@$ as category builders to a hand-crafted gold standard like WordNet.

Other researchers have likewise used WordNet as a gold standard for categorization experiments, and we replicate here the experimental set-up of Almuhareb and Poesio (2004, 2005), which is designed to measure the effectiveness of web-acquired conceptual descriptions. Almuhareb and Poesio choose 214 English nouns from 13 of WordNet’s upper-level semantic categories, and proceed to harvest property values for these concepts from the web using the Hearst-like pattern “ $a|an|the * C$ is|was”. This pattern yields a combined total of 51,045 values for all 214 nouns; these values are primarily adjectives, such as *hot* and *black* for *coffee*, but noun-modifiers of C are also allowed, such as *fruit* for *cake*. They also harvest 8934 attribute nouns, such as *temperature* and *color*, using the query “ $the * of the C$ is|was”. These values and attributes are then used as the basis of a clustering algorithm to partition the 214 nouns back into their original 13 categories. Comparing these clusters with the original WordNet-based groupings, Almuhareb and Poesio report a cluster accuracy of **71.96%** using just values like *hot* and *black* (51,045 values), an accuracy of **64.02%** using just attributes like *temperature* and *color* (8,934 attributes), and an accuracy of **85.5%** using both together (a combined 59,979 features).

How concisely and accurately does $@X$ describe a noun X for purposes of categorization? Let AP denote the set of 214 WordNet nouns used by Almuhareb and Poesio. Then $@^AP$ denotes a set of 2,209 adjectival properties; this should be contrasted with the space of 51,045 adjectival values used by Almuhareb and Poesio. Using the same clustering algorithm over this feature set, $@X$ achieves a clustering accuracy (as measured via cluster purity) of **70.2%**, compared to **71.96%** for Almuhareb and Poesio. However, when $@X$ is used to harvest a further set of attribute nouns for X , via web queries of the form “ $the P * of X$ ” (where $P \in @X$), then $@X$ augmented with this additional set of attributes (like *hands* for *surgeon*)

produces a larger space of 7,183 features. This in turn yields a cluster accuracy of **90.2%** which contrasts with Almuhareb and Poesio's **85.5%** for 59,979 features. In either case, **@X** produces comparable clustering quality to Almuhareb and Poesio, with just a small fraction of the features.

So how concisely and accurately does **?X** describe a noun *X* for purposes of categorization? While **@X** denotes a set of salient adjectives, **?X** denotes a set of comparable nouns. So this time, **?^AP** denotes a set of 8,300 nouns in total, to act as a feature space for the 214 nouns of Almuhareb and Poesio. Remember, the contents of each **?X**, and of **?^AP** overall, are determined entirely by the contents of the Google 3-grams; the elements of **?X** are not ranked in any way, and all are treated as equals. When the 8,300 features in **?^AP** are clustered into 13 categories, the resulting clusters have a purity of **93.4%** relative to WordNet. The pragmatic neighborhood of *X*, **?X**, appears to be an accurate and concise proxy for the meaning of *X*.

What about adjectives? Almuhareb and Poesio's set of 214 words does not contain adjectives, and besides, WordNet does not impose a category structure on its adjectives. In any case, the role of adjectives in the applications of section 4 is largely an affective one: if *X* is a noun, then one must have confidence that the adjectives in **@X** are consonant with our understanding of *X*, and if *P* is a property, that the adjectives in **?P** evoke much the same mood and sentiment as *P*. Our evaluation of **@X** and **?P** should thus be an affective one.

So how well do the properties in **@X** capture our sentiments about a noun *X*? Well enough to estimate the pleasantness of *X* from the adjectives in **@X**, perhaps? Whissell's (1989) dictionary of affect provides pleasantness ratings for a sizeable number of adjectives and nouns (over 8,000 words in total), allowing us to estimate the pleasantness of *X* as a weighted average of the pleasantness of each X_i in **@X** (the weights here are web frequencies for the similes that underpin **@** in section 3.2). We thus estimate the affect of all stereotype nouns for which Whissell also records a score. A two-tailed Pearson test ($p < 0.05$) shows a positive correlation of **0.5** between these estimates and the pleasantness scores assigned by Whissell. In contrast, estimates based on the pleasantness of adjectives found in corresponding WordNet glosses show a positive correlation of just **0.278**.

How well do the elements of **?P** capture our sentiments toward an adjective *P*? After all, we hypothesize that the adjectives in **?P** are highly suggestive of *P*, and vice versa. *Aristotle* and the *Jigsaw Bard* each rely on **?P** to suggest adjectives that evoke an unstated property in a metaphor or simile, or to suggest coherent blends of properties. When we estimate the pleasantness of each adjective *P* in Whissell's dictionary via the weighted mean of the pleasantness of adjectives in **?P** (again using web frequencies as weights), a two-tailed Pearson test ($p < 0.05$) shows a correlation of **0.7** between estimates and actual scores. It seems **?P** does a rather good job of capturing the feel of *P*.

6 Concluding Remarks

Creative information retrieval is not a single application, but a paradigm that allows us to conceive of many different kinds of application for creatively manipulating text. It is also a tool-kit for implementing such an application, as shown here in the cases of *Aristotle*, *Idiom Savant* and *Jigsaw Bard*.

The wildcards **@**, **?** and **^** allow users to formulate their own task-specific ontologies of ad-hoc categories. In a fully automated application, they provide developers with a simple but powerful vocabulary for describing the range and relationships of the words, phrases and ideas to be manipulated.

The **@**, **?** and **^** wildcards are just a start. We expect other aspects of figurative language to be incorporated into the framework whenever they prove robust enough for use in an IR context. In this respect, we aim to position Creative IR as an open, modular platform in which diverse results in FLP, from diverse researchers, can be meaningfully integrated. One can imagine wildcards for matching potential puns, portmanteau words and other novel forms, as well as wildcards for figurative processes like metonymy, synecdoche, hyperbolae and even irony. Ultimately, it is hoped that creative IR can serve as a textual bridge between high-level creativity and the low-level creative potentials that are implicit in a large corpus.

Acknowledgments

This work was funded in part by Science Foundation Ireland (SFI), via the Centre for Next Generation Localization. (CNGL).

References

- Almuhareb, A. and Poesio, M. (2004). Attribute-Based and Value-Based Clustering: An Evaluation. In *Proc. of EMNLP 2004*. Barcelona.
- Almuhareb, A. and Poesio, M. (2005). Concept Learning and Categorization from the Web. In *Proc. of the 27th Annual meeting of the Cognitive Science Society*.
- Barnden, J. A. (2006). Artificial Intelligence, figurative language and cognitive linguistics. In: G. Kristiansen, M. Achard, R. Dirven, and F. J. Ruiz de Mendoza Ibanez (Eds.), *Cognitive Linguistics: Current Application and Future Perspectives*, 431-459. Berlin: Mouton de Gruyter.
- Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, 11:211-227.
- Boden, M. (1994). Creativity: A Framework for Research, *Behavioural & Brain Sciences* 17(3):558-568.
- Brants, T. and Franz, A. (2006). *Web IT 5-gram Ver. 1*. Linguistic Data Consortium.
- Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13-47.
- Falkenhainer, B., Forbus, K. and Gentner, D. (1989). Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41:1-63.
- Fass, D. (1991). Met*: a method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49-90.
- Fass, D. (1997). Processing Metonymy and Metaphor. *Contemporary Studies in Cognitive Science & Technology*. New York: Ablex.
- Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge.
- Fishlov, D. (1992). Poetic and Non-Poetic Simile: Structure, Semantics, Rhetoric. *Poetics Today*, 14(1), 1-23.
- Gentner, D. (1983). Structure-mapping: A Theoretical Framework. *Cognitive Science* 7:155-170.
- Guilford, J.P. (1950) Creativity, *American Psychologist* 5(9):444-454.
- Hanks, P. (2005). Similes and Sets: The English Preposition 'like'. In: Blatná, R. and Petkevic, V. (Eds.), *Languages and Linguistics: Festschrift for Fr. Cermak*. Charles University, Prague.
- Hanks, P. (2006). Metaphoricity is gradable. In: Anatol Stefanowitsch and Stefan Th. Gries (Eds.), *Corpus-Based Approaches to Metaphor and Metonymy*, 17-35. Berlin: Mouton de Gruyter.
- Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th Int. Conf. on Computational Linguistics*, pp 539-545.
- Indurkha, B. (1992). *Metaphor and Cognition: Studies in Cognitive Systems*. Kluwer Academic Publishers, Dordrecht: The Netherlands.
- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proc. of the 17th international conference on Computational linguistics*, 768-774.
- MacCormac, E. R. (1985). *A Cognitive Theory of Metaphor*. MIT Press.
- Martin, J. H. (1990). *A Computational Model of Metaphor Interpretation*. New York: Academic Press.
- Mason, Z. J. (2004). CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System, *Computational Linguistics*, 30(1):23-44.
- Mihalcea, R. (2002). The Semantic Wildcard. In *Proc. of the LREC Workshop on Creating and Using Semantics for Information Retrieval and Filtering*. Canary Islands, Spain, May 2002.
- Navigli, R. and Velardi, P. (2003). An Analysis of Ontology-based Query Expansion Strategies. In *Proc. of the workshop on Adaptive Text Extraction and Mining (ATEM 2003)*, at ECML 2003, the 14th European Conf. on Machine Learning, 42-49
- Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York: McGraw-Hill.
- Taylor, A. (1954). Proverbial Comparisons and Similes from California. *Folklore Studies* 3. Berkeley: University of California Press.
- Van Rijsbergen, C. J. (1979). *Information Retrieval*. Oxford: Butterworth-Heinemann.
- Veale, T. (2004). The Challenge of Creative Information Retrieval. *Computational Linguistics and Intelligent Text Processing: Lecture Notes in Computer Science*, Volume 2945/2004, 457-467.
- Veale, T. (2006). Re-Representation and Creative Analogy: A Lexico-Semantic Perspective. *New Generation Computing* 24, pp 223-240.
- Veale, T. and Hao, Y. (2007). Making Lexical Ontologies Functional and Context-Sensitive. In *Proc. of the 46th Annual Meeting of the Assoc. of Computational Linguistics*.
- Veale, T. and Hao, Y. (2010). Detecting Ironic Intent in Creative Comparisons. In *Proc. of ECAI'2010, the 19th European Conference on Artificial Intelligence*.

- Veale, T. and Butnariu, C. (2010). Harvesting and Understanding On-line Neologisms. In: Onysko, A. and Michel, S. (Eds.), *Cognitive Perspectives on Word Formation*. 393-416. Mouton De Gruyter.
- Vernimb, C. (1977). Automatic Query Adjustment in Document Retrieval. *Information Processing & Management*. 13(6):339-353.
- Voorhees, E. M. (1994). Query Expansion Using Lexical-Semantic Relations. In *the proc. of SIGIR 94, the 17th International Conference on Research and Development in Information Retrieval*. Berlin: Springer-Verlag, 61-69.
- Voorhees, E. M. (1998). Using WordNet for text retrieval. *WordNet, An Electronic Lexical Database*, 285-303. The MIT Press.
- Way, E. C. (1991). Knowledge Representation and Metaphor. *Studies in Cognitive systems*. Holland: Kluwer.
- Weeds, J. and Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):433-475.
- Whissell, C. (1989). The dictionary of affect in language. R. Plutchnik & H. Kellerman (Eds.) *Emotion: Theory and research*. NY: Harcourt Brace, 113-131.
- Wilks, Y. (1978). Making Preferences More Active, *Artificial Intelligence* 11.
- Xu, J. and Croft, B. W. (1996). Query expansion using local and global document analysis. In *Proc. of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*.