Generating and Understanding Creative Comparisons

Tony Veale and **Yanfen Hao**
School of Computer Science and Informatics, University College Dublin
Dublin 4, Ireland
{Tony.Veale, Yanfen.Hao}@ucd.ie

Abstract

If creativity is a reaction against the norm, then to understand and exploit creativity one must first understand and adequately represent this norm. In other words, to produce the extraordinary, as a human or as a machine, one must first understand the ordinary. Creativity is a large theme that manifests itself in many small ways in language, and in this paper we address one such manifestation of linguistic creativity, the comparison statement. Since linguistic comparisons run the gamut from the ordinary (mundane and commonplace) to the extraordinary (i.e., novel, striking and/or humorous), they provide an excellent vehicle for understanding the interplay between norms and creativity.

Introduction

Many of the normative beliefs that one uses to reason about everyday entities and events are neither strictly true or even logically consistent. Rather, people appear to rely on a large body of folk knowledge in the form of stereotypes, clichés and other prototype-centric structures (e.g., see Lakoff, 1987). These stereotypes comprise the landmarks of our conceptual space against which other, less familiar concepts can be compared and defined. For instance, people readily employ the animal concepts Snake, Bear, Bull, Wolf, Gorilla and Shark in everyday conversation without ever having had first-hand experience of these entities. Nonetheless, our culture equips us with enough folk knowledge of these highly evocative concepts to use them as dense short-hands for all manner of behaviours and property complexes. Snakes, for example, embody the notions of treachery, slipperiness, cunning and charm (as well as a host of other, related properties) in a single, visually-charged package. To compare someone to a snake is to suggest that many of these properties are present in that person, and thus, one would well to treat that person as one would treat a real snake.

In "A Christmas Carol", Dickens (1843/1984) notes that "the wisdom of our ancestors is in the simile; and my unhallowed hands shall not disturb it, or the Country's done for" (chapter 1, page 1). In other words, stereotypical knowledge is passed down through a culture via language, most often in specific linguistic forms. The simile, as noted by Dickens, is one common vehicle for folk wisdom, one that uses explicit syntactic means (unlike metaphor; see Hanks, 2004) to mark out those concepts that are most useful as landmarks for linguistic description. Similes do not always convey truths that are universally true, or indeed, even literally true (e.g., bowling balls are not literally bald). Rather, similes hinge on properties that are possessed by prototypical or stereotypical members of a category (see Ortony, 1979), even if most members of the category do not also possess them. As a source of knowledge, similes combine received wisdom, prejudice and over-simplifying idealism in equal measure. As such, similes reveal knowledge that is pragmatically useful but of a kind that one is unlikely to ever acquire from a dictionary (or, indeed, from WordNet; see Fellbaum, 1998). Although a simpler rhetorical device than metaphor, we have much to learn about language and its underlying conceptual structure by a comprehensive study of real similes in the wild (see Roncero *et al.* 2007), not least about the recurring vehicle categories that signpost this space (see Veale and Hao, 2007).

To model our capacity for both stereotypical and creative comparison, we describe in sections 2, 3 and 4 a series of web-harvesting techniques for acquiring the largest database of simile-based comparisons of its kind. We then demonstrate the effectiveness of stereotypical representations as derived from similes in section 5. In section 6, we describe an on-line computational realization, called Aristotle, that allows users to generate and analyze creative comparisons for topics of their own choosing. Though our approach has mostly concentrated thus far on similes in English, we also demonstrate its applicability to other languages such as Chinese, and conclude the paper by reporting cross-cultural differences on the prevalence of creative irony in both languages

Acquiring Stereotypical Features from Similes

As in the study reported in Roncero *et al.* (2006), we employ the *Google* search engine as a retrieval mechanism for accessing relevant web content. However, the scale of the current exploration requires that retrieval of similes be fully automated, and this automation is facilitated both by the *Google* API and its support for the wildcard term *. In essence, we consider here only partial explicit similes conforming to the pattern "*as ADJ as a|an NOUN*", in an attempt to collect all of the salient values of ADJ for a given value of NOUN. We do not expect to identify and retrieve all similes mentioned on the world-wide-web, but to gather a large, representative sample of the most commonly used.

To do this, we first extract a list of antonymous adjectives, such as "hot" or "cold", from WordNet (Fellbaum, 1998), the intuition being that explicit similes will tend to exploit properties that occupy an exemplary point on a scale. For every adjective ADJ on this list, we send the query "*as ADJ as* *" to Google and scan the first 200 snippets returned for different noun values for the wildcard *. From each set of snippets we can ascertain the relative frequencies of different noun values for ADJ. The complete set of nouns extracted in this way is then used to drive a second phase of the search. In this phase, the query "*as * as a NOUN*" is used to collect similes that may have lain beyond the 200-snippet horizon of the original search, or that hinge on adjectives not included on the original list. Together, both phases collect a wide-ranging series of core samples (of 200 hits each) from across the web, yielding a set of 74,704 simile instances (of 42,618 unique types) relating 3769 different adjectives to 9286 different nouns

Simile Annotation

Many of these similes are not sufficiently well-formed for our purposes. In some cases, the noun value forms part of a larger noun phrase: it may be the modifier of a compound noun (as in "bread lover"), or the head of complex noun phrase (such as "gang of thieves"). In the former case, the compound is used if it corresponds to a compound term in WordNet and thus constitutes a single lexical unit; if not, or if the latter case, the simile is rejected. Other similes are simply too contextual or under-specified to function well in a null context, so if one must read the original document to make sense of the simile, it is rejected. More surprisingly, perhaps, a substantial number of the retrieved similes are ironic, in which the literal meaning of the simile is contrary to the meaning dictated by common sense. For instance, "as hairy as a bowling ball" (found once) is an ironic way of saying "as hair<u>less</u> as a bowling ball" (also found just once). Many ironies can only be recognized using world (as opposed to word) knowledge, such as "as sober as a Kennedy" and "as tanned as an Irishman". In addition, some similes hinge on a new, humorous sense of the adjective, as in "as fruitless as a butcher-shop" (since the latter contains no fruits) and "as pointless as a beach-ball" (since the latter has no points).

Given the creativity involved in these constructions, one cannot imagine a reliable automatic filter to safely identify bona-fide similes. For this reason, the filtering task was performed by human judges, who annotated 30,991 of these simile instances (for 12,259 unique adjective/noun pairings) as non-ironic and meaningful in a null context; these similes relate a set of 2635 adjectives to a set of 4061 different nouns. In addition, the judges also annotated 4685 simile instances (of 2798 types) as ironic; these similes relate 936 adjectives to a set of 1417 nouns. Perhaps surprisingly, ironic pairings account for over 13% of all annotated simile instances and over 20% of all annotated simile types .

Harvesting Creatively Enhanced Comparisons

So much for the ordinary. Such stereotypes can be used for creative purposes when it is clear to an observer/listener that something ordinary is being exploited in a novel way, that is, when the observer is made aware of the extraordinary in the ordinary. For instance, consider the following clichéd comparisons for something "ugly" (the numbers in parentheses are frequency-counts for the corresponding web similes):

{*toad(11), wart(5), dog(4), witch(3), warthog(3), toad(3), hippopotamus(3), dump(2), mule(2), blister(2), baboon(2), crab(2), ...*}

Now consider how these commonplace descriptions might be enhanced and extended by the simple addition of a single adjectival modifier:

{*squashed:toad(2), genital:wart(2), old:toad(1), horned:toad(1), dirty:pig(1), busted:blister(1), one-eared:dog(1), chained:dog(1), rabid:dog(1), shaved:mule(1)*}

Though some are more creative than others, these extended comparisons are each creative to the extent that each achieves a wonderfully picturesque concision. That is, a single modifier is used to add a wealth of affect-rich information: if a toad is a model of ugliness, a "squashed toad" evokes a far uglier image, while also interjecting an element of pathos and humour. Though the images reside in the realm of the possible, the humour often arises from a sense of incongruity (e.g., see Ritchie, 1999), as in "shaved mule" (which begs the question "why would one ever shave a mule?").

Enhanced descriptions such as these can also be harvested from the web, using the Google query "as ADJ as a|an * NOUN", where ADJ and NOUN correspond to the elements of a previously acquired simile (such as *ugly*:*toad*), and adjectival values are sought for the * wildcard. Using the filtered simile set from section 2.1. as a retrieval basis, this query allows us to retrieve a further 5729 enhanced similes such as those listed above for "ugly". Since these are elaborated forms of previously validated similes, the degree of noise and irony found in these enhanced similes is negligible: just 108 enhanced similes (less than 2%) are found to be ironic, while just 72 (just over 1%) are dismissed outright as noise. We note that because of this rarity, and because the base similes from which these 108 ironies are derived are themselves highly stereotypical, the resulting ironies – such as "as accurate as a blind archer" and "as precise as a drunken surgeon" – exhibit a subversive form of humorous creativity. One can also see from these cases how templates for generating other ironic comparisons, such as "as accurate as a blind X" (where X is a stereotype for accuracy), can be induced automatically, though this is a topic we leave for future research.

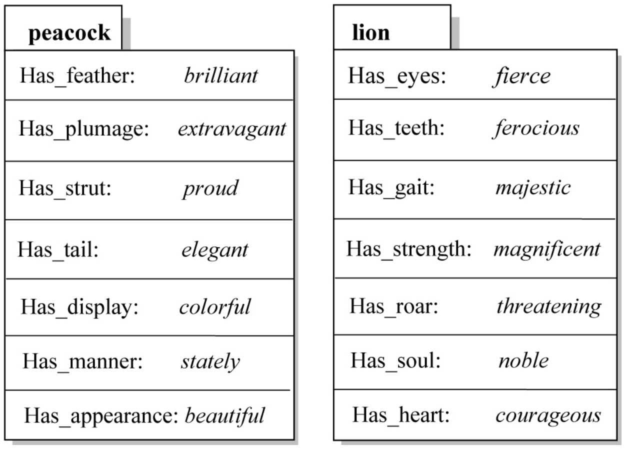
Building Stereotypical Frame Representations

Each bona-fide simile contributes a different salient property to the representation of a vehicle concept. In our data, one half (49%) of all bona-fide vehicle nouns occur in two or more similes, while one third occur in three or more and one fifth occur in four or more. The most frequently used figurative vehicles can have many more; “snowflake”, for instance, is ascribed over 30 in our database, including: *white, pure, fresh, beautiful, natural, delicate, intricate, delicate, identifiable, fragile, light, dainty, frail, weak, sweet, precious, quiet, cold, soft, clean, detailed, fleeting, unique, singular, distinctive* and *lacy*.

Because the same adjectival properties are associated with multiple vehicles, the resulting property graph allows different vehicles to be perceived as similar by virtue of these shared properties. For instance, Ninja and Mime are deemed similar by virtue of the shared property *silent*, while Artist and Surgeon are similar by virtue of the properties *skilled*, *sensitive* and *delicate*. Nonetheless, it can be claimed the property level is simply too shallow to allow for nuanced similarity judgements. For instance, are ninjas and mimes silent in the same way? Both surgeons and bloodhounds are prototypes of sensitivity, but the former has sensitive *hands* while the latter has a sensitive *nose*. To put these properties in context, we need to know the specific facet of each concept that is modified, so that sensible comparisons can be made. In effect, we need to move from a simple property-ascription representation to a richer, *frame:slot:filler* representation. In such a scheme, the property sensitive is a typical filler for the hands slot of Surgeon and the nose slot of Bloodhound, thereby disallowing any mis-matched comparisons.

This process of frame construction can also be largely automated via targeted web-search. For every bona-fide simile-type “as ADJ as a Noun$_{vehicle}$”, we automatically generate the web-query “the ADJ * of a Noun$_{vehicle}$” and harvest the top 200 results from Google. From these snippets, we then extract all noun values of the wildcard *. In many cases, these noun values are precisely the conceptual facets we desire for a culturally-accurate and nuanced representation, ranging from *hands* for Surgeon to *roar* for Lion to *eye* for Hawk. The frequency of these values also allows us to create a textured representation for each concept, so that e.g., both *hands* and *eye* are notable facets for surgeon, but the latter is higher ranked. However, this web-pattern also yields a non-trivial amount of noise: while “the proud strut of a peacock” is very revealing about the concept Peacock, the snippet “the proud owner of a peacock” is not. Quite simply, we seek to fill intrinsic facets of a concept like *hands*, *eye*, *gait* and *strut* that contribute to the folk definition of the concept, while ignoring extrinsic and contingent facets such as *owner*, *husband*, *brother* and so on.

One can look to specific abstractions in WordNet – such as {*trait*} - to serve as a filter on the facet-nouns that are extracted, but such a simple filter would be unduly coarse. Instead, we consider all facet-nouns, but generalize the WordNet vehicle-senses to which they are attached, to create a high-level mapping of vehicle types (such as Person, Animal, Implement, Substance, etc.) to facets (such as *hands, eye, sparkle, father*, etc.). This high-level (and considerably more compressed) map is then human-edited, to remove any facets that are unrevealing or simply appropriate for the WordNet vehicle type. In this editing process (which requires about one man-day), contingent facets such as *father, wife*, etc. are quickly identified and removed. As can be seen in the examples of Lion and Peacock in Figure 1, the slot:filler pairs that are acquired for each concept do indeed reflect the most relevant cultural associations for these concepts. Moreover, there is a great deal of anthropomorphic rationalization of an almost poetic nature about these representations, of the kind that is instantly recognizable to native speakers of a language but which one would be hard pressed to find in a conventional dictionary (except insofar as some lexical concepts may give rise to additional word senses, such as ”peacock” for a proud and flashily dressed person).



peacock	
Has_feather:	*brilliant*
Has_plumage:	*extravagant*
Has_strut:	*proud*
Has_tail:	*elegant*
Has_display:	*colorful*
Has_manner:	*stately*
Has_appearance:	*beautiful*

lion	
Has_eyes:	*fierce*
Has_teeth:	*ferocious*
Has_gait:	*majestic*
Has_strength:	*magnificent*
Has_roar:	*threatening*
Has_soul:	*noble*
Has_heart:	*courageous*

Figure 1: Frame:slot:filler stereotype structures for Peacock and Lion.

Overall, frame representations of this kind are acquired for 2218 different WordNet noun senses, yielding a combined total of 16,960 slot:filler pairings (or an average of 8 slot:filler pairs per frame). As the examples of Figure 1 demonstrate, these frames provide a level of representational finesse that greatly enriches the basic property descriptions yielded by similes alone. To answer an earlier question then, mimes and ninjas are now similar by virtue of each possessing the slot:filler *Has_silent:art*. But as this and other examples suggest, the introduction of finely discriminating frame structures can decrease a system’s ability to recognize similarity, if comparable slots or fillers are given different names. In Figure 1, for instance, a human can easily recognize that *Has_strut:proud* and *Has_gait:majestic* are similar properties, but to a computer they can appear very different ideas. WordNet can play a significant role in reconciling these superficial differences in structure (e.g., by recognizing the obvious relationship between *strut* and *gait*), while corpus-based co-occurrence models can reveal the comparable nature of *proud* and *majestic*. This work, however, is outside the scope of the current paper and is the subject of future development and research.

Approach	*accuracy*	*features*
Almuhareb + Poesio	71.96%	51,045
Simile-derived stereotypes	70.2%	2,209

Table 1: Results for experiment 1 (214 nouns, 13 WN categories).

Empirical Evaluation

Stereotypes persist in language and culture because they are, more often than not, cognitively useful: by emphasizing the most salient aspects of a concept, a stereotype acts as a dense conceptual description that is easily communicated, widely shared, and which supports rapid inference. To demonstrate the usefulness of stereotype-based concept descriptions, we replicate here the clustering experiments of Almuhareb and Poesio (2004,2005), who in turn demonstrated that conceptual features that are mined from specific textual patterns can be used to construct WordNet-like ontological structures.

Almuhareb and Poesio (2004) used as their experimental basis a sampling of 214 English nouns from 13 of WordNet's upper-level semantic categories, and proceeded to harvest adjectival features for these noun-concepts from the web using the textual pattern "**[a** | **an** | **the]** * **C [is** | **was]**". This pattern yielded a combined total of 51,045 value features for these 214 nouns, such as *hot, black*, etc., which were then used as the basis of a clustering algorithm in an attempt to reconstruct the WordNet classifications for all 214 nouns. Clustering was performed by the CLUTO-2.1 package (Karypis, 2003), which partitioned the 214 nouns in 13 categories on the basis of their 51,045 web-derived features. Comparing these clusters with the original WordNet-based groupings, Almuhareb and Poesio report a clustering accuracy of 71.96%. In a second, larger experiment, Almuhareb and Poesio (2005) sampled 402 nouns from 21 different semantic classes in WordNet, and harvested 94,989 feature values from the web using the same textual pattern. They then applied the repeated bisections clustering algorithm to this larger data set, and report an initial cluster purity measure of 56.7%. Suspecting that a noisy feature set had contributed to the apparent drop in performance, these authors then proceed to apply a variety of noise filters to reduce the set of feature values to 51,345, which in turn leads to an improved cluster purity measure of 62.7%.

We replicated both of Almuhareb and Poesio's experiments on the same experimental data-sets (of 214 and 402 nouns respectively), using instead the English simile pattern "as * as a NOUN" to harvest features for these nouns from the web. Note that in keeping with the original experiments, no hand-tagging or filtering of these features is performed, so that every raw match with the simile pattern is used. Overall, we harvest just 2209 feature values for the 214 nouns of experiment 1, and 5547 features for the 402 nouns of experiment 2. A comparison of both sets of results for experiment 1 is shown is Table 1, while a comparison based on experiment 2 is shown is Table 2.

Approach	*Cluster purity*	*Cluster entropy*	*features*
Almu. + Poesio (no filtering)	56.7%	38.4%	94,989
Almu. + Poesio (with filtering)	62.7%	33.8%	51345
Simile-derived stereotypes (no filtering)	64.3%	33%	5,547

Table 2: Results for experiment 2 (402 nouns, 21 WN categories).

Concluding Remarks

In this paper we have presented an approach to acquiring the stereotypical cultural associations that pervade our everyday use of language yet which one rarely finds in authoritative linguistic resources like dictionaries and encyclopaedias. Our means of acquiring these associations – via explicit similes that are mined from the internet – has several important consequences for a model of creative comparison. First, we acquire associations that are neither necessarily true or necessarily consistent with each other, but which people happily assume to be true and consistent for purposes of habitual reasoning. Second, a large-scale mining effort allows us to identify the most frequently used vehicles of comparison, and thus, the most salient landmarks of our shared conceptual space. Thirdly, we identify the most salient properties of these landmarks, also frequency weighted, as well as the most notable conceptual facets of these landmarks. These combinations of facets and properties (i.e., slot:filler pairings) have a poetic quality that can be used to drive the automatic natural-language generation of creative descriptions.

Aristotle: A Web-based Description Generator

The approach described in this paper is embodied in a computational realization called "Aristotle", accessible on-line of the *http://afflatus.ucd.ie/aristotle/* web-site. Aristotle generates descriptions, both stereotypical and creative, by prompting a user to enter a topic to be described, as well as an adjectival property of that topic to highlight. For instance, suppose one chooses the topic "supermodel" and the property "skinny"; Aristotle uses WordNet (Fellbaum, 1998) to generalize "supermodel" to the generic category "person", and then uses its database of similes to generate the following list of potential comparisons for a person:

{*snake(58), noodle(57), post(46), miser(50), twig(49), stick(47), pencil(41), rope(34), mosquito(26), scarecrow(23), thread(17), cadaver(17)*}

The number given in parentheses for each comparison concept C specifies the web-frequency (as obtained from Google) for the corresponding phrase "C-like person"; thus, *snake(58)* denotes the fact that "snake-like person" occurs 58 times on the web. Now, looking also to the enhanced simile set (as acquired in section 3), we find the following extensions of "skinny" comparisons:

{*old:rail(1), little:rail(1), anorexic:whippet(1)*}

Note that the phrase "whippet-like person" cannot be found on the web, so Aristotle does not initially consider "whippet" a suitable stand-alone comparison for "supermodel". Nonetheless, "anorexic person" does occur 474 times (and "anorexic supermodel" 148 times), so the extended comparison "like an anorexic whippet" is considered meaningful and valid for "supermodel".

Other Languages, Other Cultures

Despite these benefits, our continued reference to the notion of "culture" may seem misplaced given our focus on English-language similes and on English-centric frame representations. Nonetheless, we see this work as a platform from which to explore the cultural diversity of creative descriptions and their underlying representational basis, and to this end, we are currently in the process of replicating the simile-based approach for Chinese and Korean. In the case of Chinese, we are using the Chinese-English bilingual ontology of HowNet (Dong and Dong, 2006) to drive the acquisition of Chinese web similes, and from these, to obtain Chinese-centric stereotype representations To see how similes reflect different biases in different cultures, consider that of the 12,259 unique adjective/noun pairings judged as bona-fide (non-ironic) in section 2.1, only 2,440 (or 20%) have a Chinese translation that can also be found on the web (where translation is performed using the bilingual HowNet). The replication rate for the ironic similes of section 2.1. is even lower, at 5%, reflecting the fact that ironic comparisons are more creatively ad-hoc and less culturally entrenched than non-ironic similes. Our current efforts toward the mining of Chinese web-texts are yielding a set of similes – and thus conceptual descriptions (both properties and frames) – that are substantially different from the English-language set described here, enabling us to generate Chinese descriptions that are creative but culturally appropriate.

References

Almuhareb, A. and Poesio.M. 2004. Attribute-Based and Value-Based Clustering: An Evaluation. In *Proceedings of EMNLP 2004*, 158–165. Barcelona, Spain.

Almuhareb, A. and Poesio.M. 2005. Concept Learning and Categorization from the Web. In *Proceedings of CogSci 2005, the 27th Annual Conference of the Cognitive Science Society*, New Jersey: Lawrence Erlbaum.

Dickens, C. (1843/1984). A Christmas Carol.: Puffin Books, Middlesex, UK.

Dong, Z and Dong, Q. 2006. *HowNet and the Computation of Meaning*. World Scientific: Singapore.

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

Hanks, P. 2004. The syntagmatics of metaphor. *International Journal of Lexicography*, 17(3).

Karypis, G. 2003. CLUTO: A clustering toolkit. University of Minnesota.

Lakoff, G. 1987. *Women, fire and dangerous things*. Chicago University Press.

Ortony, A. 1979. Beyond literal similarity. *Psychological Review*, 86, 161–180.

Rawson, H. 1995. A Dictionary of Euphemisms and Other Doublespeak, New York: Crown Publishers.

Ritchie, G. 1999. Developing the incongruity-resolution theory. In *proceedings of the AISB Symposium on Creative Language*. Edinburgh, Scotland, 78 – 85.

Roncero, C., Kennedy, J. M., and Smyth, R. 2006. Similes on the internet have explanations. *Psychonomic Bulletin and Review*, 13(1), 74–77.

Veale, T. and Hao, Y. 2007. Making Lexical Ontologies Functional and Context-Sensitive. In *proceedings of ACL 2007, the 45th Annual Meeting of the Association of Computational Linguistics*, 57–64. Prague, Czech Republic.