Enriching WordNet with Folk Knowledge and Stereotypes

Tony Veale[1] and Yanfen Hao[1]

[1] School of Computer Science and Informatics, University College Dublin, Dublin, Ireland
{Tony.Veale, Yanfen.Hao}@UCD.ie

Abstract. The knowledge that is needed to understand everyday language is not necessarily the knowledge one finds in an encyclopedia or dictionary. Much of this is "folk" knowledge, based on stereotypes and culturally-inherited associations that do not hold in all situations, or which may, strictly speaking, be false. We can open a linguistic window onto this knowledge through simile, since explicit similes make use of highly evocative and inference-rich concepts to ground comparisons and make the unfamiliar seem familiar. This paper describes a means of enriching WordNet with commonly ascribed cultural properties by mining explicit similes of the form "as ADJ as a NOUN" from the internet. We also show how these properties can be leveraged, through further web search, into rich frame structures for the most evocative WordNet concepts.

Keywords: simile, folk knowledge, frame representation.

1 Introduction

Many of the 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 prototypes 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 do well to treat that person as one would treat a real snake.

Descriptors like "snake", "shark" and "wolf" find a great deal of traction in everyday conversation because they are "dense descriptors" – they convey a great deal of useful information in a simple and concise way. The information imparted is open-ended, so that a listener may take meaning X from the description when it is initially used (e.g., that a given person is treacherous) and meaning X+Y (e.g., that

this person is both treacherous *and* charming) in a later, more informed context. But the information imparted is rarely of the kind one finds in a dictionary or encyclopaedia, or in a resource like WordNet (Fellbaum, 1998), because it is neither contributes to the definition of the given concept or is actually true of that concept. Insofar as WordNet is used to make sense of real texts by real, culturally-grounded speakers, it can be enriched considerably by the addition of such stereotypical knowledge. But where can this knowledge be found and exploited?

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, folk 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). 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).

In this paper we describe a means through which we can enrich WordNet with stereotypical folk-knowledge from similes that are mined from the text of the world-wide web. We describe the Google-based mining process in section 2, before describing how the acquired knowledge is sense-linked to WordNet in section 3. In section 4 we describe on-going work to elaborate this property-rich knowledge into more complex frame-representations, before providing an empirical evaluation of the basic properties in section 5. The paper concludes with thoughts on future work in section 6.

2 Acquiring Knowledge from Simile

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.

2.1 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 hairless 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.

3 Establishing Links to WordNet

It is important to know which sense of a noun is described by a simile if an accurate conceptual picture is to be constructed. For instance, "as stiff as a zombie" might refer either to a reanimated corpse or to an alcoholic cocktail (both are senses of "zombie" in WordNet, and drinks can be "stiff" too). Sense disambiguation is especially important if we hope to derive meaningful correlations from property co-occurrences; for instance, zombies are described in web similes as exemplars of not just stiffness, but of coldness, slowness and emotionlessness. If such co-occurrences are observed often enough, a cognitive agent might usefully infer a causal relationship among pairs of properties.

Disambiguation is trivial for nouns with just a single sense in WordNet. For nouns with two or more fine-grained senses that are all taxonomically close, such as "gladiator" (two senses: a boxer and a combatant), we consider each sense to be a suitable target. In some cases, the WordNet gloss for as particular sense will actually mention the adjective of the simile, and so this sense is chosen. In all other cases, we employ a strategy of mutual disambiguation to relate the noun vehicle in each simile to a specific sense in WordNet. Two similes "as ADJ as $NOUN_1$" and "as ADJ as $NOUN_2$" are mutually disambiguating if $NOUN_1$and $NOUN_2$ are synonyms in WordNet, or if some sense of $NOUN_1$ is a hypernym or hyponym of some sense of $NOUN_2$ in WordNet. For instance, the adjective "scary" is used to describe both the noun "rattler" and the noun "rattlesnake" in bona-fide (non-ironic) similes; since these nouns share a sense, we can assume that the intended sense of "rattler" is that of a dangerous snake rather than a child's toy. Similarly, the adjective "brittle" is used to describe both saltines and crackers, suggesting that it is the bread sense of "cracker" rather than the hacker, firework or hillbilly senses (all in WordNet) that is intended.

These heuristics allow us to automatically disambiguate 10,378 bona-fide simile types (85% of those annotated), yielding a mapping of 2124 adjectives to 3778 different WordNet senses. Likewise, 77% (or 2164) of the simile types annotated as ironic are disambiguated automatically. A remarkable stability is observed in the alignment of noun vehicles to WordNet senses: 100% of the ironic vehicles always denote the same sense, no matter the adjective involved, while 96% of bona-fide vehicles always denote the same sense. This stability suggests two conclusions: the disambiguation process is consistent and accurate; but more intriguingly, only one coarse-grained sense of any word is likely to be sufficiently exemplary of some property to be useful as a simile vehicle.

4 Acquiring 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, 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 $\text{Noun}_{\text{vehicle}}$" (all 10,378 of them that have been WordNet-linked in section 3), we automatically generate the web-query "the ADJ * of a $\text{Noun}_{\text{vehicle}}$" 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.

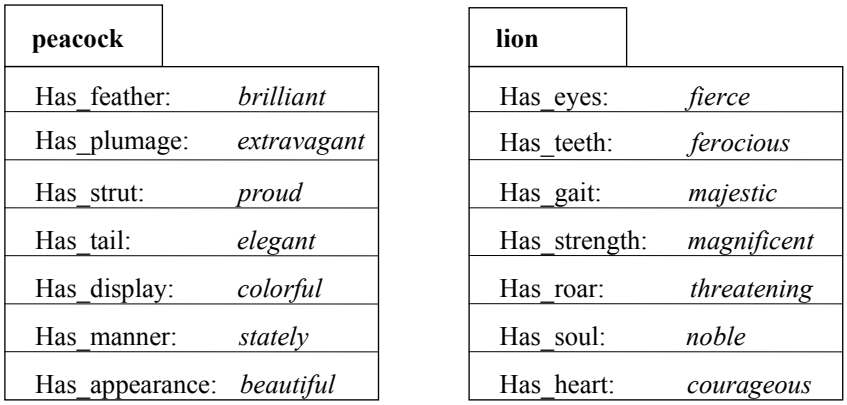

Fig. 1. The acquired Frame:slot:filler representations for Peacock and Lion.

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).

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_ art: silent*. 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.

5 Empirical Evaluation

If similes are indeed a good place to mine the most salient properties of WordNet's lexical concepts, we should expect the set of properties for each concept to accurately

predict how that concept is perceived as a whole. For instance, humans – unlike computers – do not generally adopt a dispassionate view of ideas, but rather tend to associate certain positive or negative feelings, or affective values, with particular ideas. Unsavoury activities, people and substances generally possess a negative affect, while pleasant activities and people possess a positive affect. Whissell (1989) uses human-assigned ratings to reduce the notion of affect to a single numeric dimension, to produce a *dictionary of affect* that associates a numeric value in the range 1.0 (most unpleasant) to 3.0 (most pleasant) with over 8000 words across a range of syntactic categories (including adjectives, verbs and nouns). So to the extent that the adjectival properties yielded by processing similes paint an accurate picture of each noun vehicle, we should be able to predict the affective rating of each vehicle via a weighted average of the affective ratings of the adjectival properties ascribed to these vehicles (i.e., where the affect of each adjective contributes to the estimated affect of a noun in proportion to its frequency of co-occurrence with that noun in our web-derived simile data). More specifically, we should expect ratings estimated via these simile-derived properties to exhibit a strong correlation with the independent ratings of Whissell's dictionary.

To determine whether similes do offer the clearest perspective on a concept's most salient properties, we calculate and compare this correlation using the following data sets:

A. Adjectives derived from annotated bona-fide (non-ironic) similes of section 2.1.

B. Adjectives derived from all annotated similes (both ironic and non-ironic).

C. Adjectives derived from ironic similes only.

D. All adjectives used to modify the given vehicle noun in a large corpus. We use over 2-gigabytes of text from the online encyclopaedia Wikipedia as our corpus.

E. All adjectives used to describe the given vehicle noun in any of the WordNet text glosses for that noun. For instance, WordNet defines Espresso as "strong black coffee made …" so this gloss yields the properties strong and black for Espresso.

Predictions of affective rating were made from each of these data sources and then correlated with the ratings reported in Whissell's dictionary of affect using a two-tailed Pearson test ($p < 0.01$). As expected, property sets derived from bona-fide similes only (A) yielded the best correlation (+0.514) while properties derived from ironic similes only (C) yielded the worst (-0.243); a middling correlation coefficient of 0.347 was found for all similes together, demonstrating the fact that bona-fide similes outnumber ironic similes by a ratio of 4 to 1. A weaker correlation of 0.15 was found using the corpus-derived adjectival modifiers for each noun (D); while this data provides far richer property sets for each noun vehicle (e.g., far richer than those offered by the simile database), these properties merely reflect potential rather than intrinsic properties of each noun and so do not reveal what is most salient about a vehicle concept. More surprisingly, perhaps, property sets derived from WordNet

glosses (E) are also poorly predictive, yielding a correlation with Whissell's affect ratings of just 0.278.

While it is true that the WordNet-derived properties in (E) are not sense-specific, so that properties from all senses of a noun are conflated into a single property set for that noun, this should not have dramatic effects on predictions of affective rating. Instead, if one sense of a word acquires a negative connotation, then following what is often called "Gresham's law of language" (Rawson, 1995), the "bad meanings should drive out the good" so that the word as a whole becomes tainted. Rather, it may be that the adjectival properties used to form noun definitions in WordNet are simply not the most salient properties of those nouns. To test this hypothesis, we conducted a second experiment wherein we automatically generated similes for each of the 63,935 unique adjective-noun associations extracted from WordNet glosses, e.g., "as strong as espresso", "as Swiss as Emmenthal" and "as lively as a Tarantella", and counted how many of these manufactured similes can be found on the web, again using *Google*'s API.

We find that only 3.6% of these artificial similes have attested uses on the web. From this meagre result we can conclude that: a) few nouns are considered sufficiently exemplary of some property to serve as a meaningful vehicle in a figure of speech; b) the properties used to describe concepts in the glosses of general purpose resources like WordNet are not always the properties that best reflect how humans actually think about, and use, these concepts. Of course, the truth is most likely to lie somewhere between these two alternatives. The space of potential similes is doubtless much larger than that currently found on the web, and many of the similes generated from WordNet are probably quite meaningful and apt. However, even WordNet-based similes that can be found on the web are of a different character to those that populate our database of annotated web-similes, and only 9% of the web-attested WordNet similes (or 0.32% overall) also reside in this database. Thus, most (> 90%) of the web-attested WordNet similes must lie outside the 200-hit horizon of the acquisition process described in section 2, and so are less frequent (or used in less authoritative pages) than our acquired similes.

6 Conclusion

In this paper we have presented an approach to enriching WordNet with the 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 our enrichment scheme. 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 landmarks of our shared conceptual space that are most deserving of enrichment in WordNet. Thirdly, we identify the most salient properties of these landmarks, also frequency weighted, as well as the most notable conceptual facets of these landmarks. Interestingly, these

combinations of facets and properties (i.e., slot:filler pairings) have a poetic quality that can, in future work, be exploited in the automatic natural-language generation of creative descriptions.

Despite these benefits, our continued reference to the notion of "culture" may seem misplaced given our focus on English-language similes and an English-language WordNet. Nonetheless, we see this work as a platform from which to explore the cultural diversity of ontological categorizations, and to this end, we are currently planning to replicate this approach for Chinese and Korean. In the case of Chinese, we intend the enrichment process to apply to the Chinese-English lexical ontology of HowNet (Dong and Dong, 2006). 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. We can thus expect that the mining of Chinese texts on the web will yield a set of similes – and thus conceptual descriptions (both properties and frames) – that substantially differs from the English-language set described here, to enrich HowNet in an altogether different, culturally-specific way.

References

1. Dickens, C. A Christmas Carol.: Puffin Books, Middlesex, UK (1843/1984)
2. Dong, Z and Dong, Q. HowNet and the Computation of Meaning. World Scientific: Singapore (2006)
3. Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA (1998)
4. Hanks, P. The syntagmatics of metaphor. Int. Journal of Lexicography, 17(3). (2004)
5. Lakoff, G. Women, fire and dangerous things. Chicago University Press (1987)
6. Ortony, A. Beyond literal similarity. Psychological Review, 86, pp. 161--180. (1979)
7. Rawson, H. A Dictionary of Euphemisms and Other Doublespeak, New York: Crown Publishers (1995)
8. Roncero, C., Kennedy, J. M., and Smyth, R. Similes on the internet have explanations. Psychonomic Bulletin and Review, 13(1), pp. 74--77. (2006)
9. Veale, T. and Hao, Y. Making Lexical Ontologies Functional and Context-Sensitive. In proceedings of ACL 2007, the 45th Annual Meeting of the Association of Computational Linguistics, pp. 57--64. Prague, Czech Republic (2007)
10. Whissell, C. The dictionary of affect in language. In R. Plutchnik & H. Kellerman (Eds.) (1989)
11. Emotion: Theory and research. New York, Harcourt Brace, pp. 113--131.