Teaching WordNet to *Sing like an Angel* and *Cry like a Baby*:
Learning Affective Stereotypical Behaviors from the Web

Tony Veale
Division of Web Science & Technology
Korea Advanced Institute of
Science and Technology (KAIST)
tony.veale@gmail.com

Abstract

Just as words have the potential to mean different things in different contexts, so too can their affective intent vary from one context to another. Thus, in some contexts one might feel complimented to be described as *cunning*, but feel aggrieved and insulted to be so described in another. As concepts become more complex and multifaceted, and accrete more layers of stereotypical associations, their ability to assume different affective profiles in different contexts also increases. Complimentary uses of the stereotype *baby*, for instance, will emphasize the positive behaviors of babies, while insulting uses have many negative behaviors to draw upon and accentuate. In this paper we propose a two-level organization for the affective lexicon, one that can work well with different WordNets. At the first level, stereotype-denoting terms are associated with a rich and nuanced description of their potential behaviors; at the second level, these behaviors are mapped to their affect profiles. This current paper focuses primarily on the first level.

1 Introduction

Context exerts a powerful yet largely unseen influence on our interpretation of natural language utterances. It is context that primes our expectations, to focus our attention on just those senses of a word that are relevant to its linguistic and pragmatic setting. In this way, context successfully hides from us the true complexity of the words we use, and so it can be a surprising experience to open a dictionary, or browse a WordNet (Fellbaum, 1998), and see just how many different meanings and nuances a word can convey. The same can be said of a word's *affect*: in context, a word seems to mean just what it is intended to mean, and carry just the right emotional overtones and mood. But viewed out of context, the mapping of words to affect is never quite so direct. Just as words can have many senses, so too can they have a multiplicity of affective uses.

The sense inventories that lexicographers compile for a polysemous word offer a good approximation of the word's potential to convey meaning, but affect can operate across sense boundaries and even within individual senses, at the sub-sense level. Consider the word "baby", used to denote a human infant. In some contexts the word carries a positive affect: babies can be cute and adorable, curious and trusting, and an obvious target of love and affection, especially when asleep. Crying babies, however, can be selfish, whining, drooling, hissing, tantrum-throwing little monsters. Both views are stereotypical of human babies, and either can be intended when a speaker uses the term "baby" figuratively, whether to describe a beloved partner or an annoying colleague. This is a matter of conceptual perspective, not of lexical sense, and many other words exhibit a similar affective duality; "teenager" for instance can mean "whining brat" just as easily as "growing adolescent". The concepts *Baby* and *Teenager* are complex and multifaceted, and different uses in context may highlight different stereotypical behaviors of each. Their affective meaning in context is therefore not so much a function of which lexical sense is intended but of which behaviors are highlighted, and of the perceived affect of those behaviors.

Before we can build a nuanced model of affect for a lexical resource like WordNet, we first need to understand the stereotypical behaviors on which affect is determined. With a sufficiently

rich behavioral model, we can determine the affect of a word like "baby" or "teenager" on a case-by-case and context-by-context basis, rather than wiring a one-size-fits-all measure of average affect directly into the lexicon. In short, we propose a two-level structure for a context-sensitive affective lexicon: a mapping of word-concepts to their normative stereotypical behaviors (e.g. *mewling*, *shrieking*, *drooling*, *sleeping* and *smiling*); and an affective profile of those behaviors (e.g. indicating the degree to which *shrieking* is unpleasant and *smiling* is pleasant). The affect of a word/concept in context can then be calculated as a function of the affect of its stereotypical behaviors that are primed in that context.

In this paper we focus on the first stage of the model – the construction of a rich behavior-net that associates stereotypical concepts with their expected behaviors. This stage will serve as the foundation for a subsequent model of context-sensitive lexical affect. We start in section 2 with a survey of related work in the area of stereotypes and affect, before outlining our current approach in section 3. We report on the scale of this work, and its current state, in section 4, and conclude the paper in section 5 with a brief preview of the next stage of construction for our behavior-based model of lexical affect.

2 Related Work

In its simplest form, an affect lexicon assigns an affective score – along one or more dimensions – to each word or sense. Whissell's (1989) *Dictionary of Affect*, for instance, assigns a trio of numeric scores to each of its 8000+ words to describe three psycholinguistic dimensions: *pleasantness*, *activation* and *imagery*. In the DoA, the lowest pleasantness score of 1.0 is assigned to words like *abnormal* and *ugly*, while the highest, 3.0, is assigned to words like *wedding* and *winning*. Less extreme words are assigned pleasantness scores closer to the DoA mean of 1.84. Whissell's DoA is based on human ratings, but Turney (2002) shows how such scores can be assigned automatically, using statistical measures of word association in web text.

For reliable results on a large-scale, Mohammad & Turney (2010) used the Mechanical Turk to elicit human ratings of the emotional content of different words. Ratings were sought along the eight primary emotional dimensions identified by Plutchik (1980): *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*. Automated tests were used to exclude unsuitable raters, and in all, 24,000+ word-sense pairs were annotated by five different raters. Thus, words that suggest fearful contexts, like *threat*, *hunter* and *acrobat*, are all assigned a significant score on the *fear* dimension, while *disease* and *rat* score highly on the *disgust* dimension.

Strapparava & Valitutti (2004) provide a set of affective annotations for a subset of WordNet's synsets in a resource called *Wordnet-affect*. The annotation labels, called *a-labels*, focus on the cognitive dynamics of emotion, allowing one to distinguish e.g. between words that denote an *emotion-eliciting situation* and those than denote an *emotional response*. Esuli & Sebastiani (2006) also build directly on WordNet as their lexical platform, using a semi-supervised learning algorithm to assign a trio of numbers – *positivity*, *negativity* and *neutrality* – to word senses in their newly derived resource, *SentiWordNet*. (Note that *Wordnet-affect* also supports these three dimensions as a-labels, as well as a fourth, *ambiguous*). Esuli & Sebastiani (2007) improve on their affect scores by running a variant of the PageRank algorithm (see Mihalcea and Tarau, 2004) on the implicit graph structure that tacitly connects word-senses in WordNet via their textual glosses.

These lexicons attempt to capture the affective profile of a word/sense when it is used in its most normative and stereotypical guise, but they do so without an explicit model of stereotypical meaning. Veale & Hao (2007) describe a web-based approach to acquiring such a model. They note that since the simile pattern "as ADJ as DET NOUN" presupposes that NOUN is an exemplar of ADJ*ness*, it follows that ADJ must be a highly salient property of NOUN. Veale and Hao thus harvested tens of thousands of instances of this pattern from the web, to extract sets of properties (ADJ) for thousands of commonplace terms (NOUN). They show if one estimates the pleasantness of a term like *snake* or *artist* as a weighted average of the pleasantness of its properties (like *sneaky* or *creative*) in a resource like Whissell's DoA, the estimated scores show a reliable correlation with the DoA's own scores. In other words, it makes computational sense to calculate the affect of a word-concept as a function of the affect of its most salient semantic properties.

These differing approaches are reconciled in the two-level model outlined here. A variant of the approach in Veale & Hao (2007) is used to acquire a model of stereotypical behaviors from the web. The affective profile of these behaviors

can be described using any of the above approaches, such as DoA or SentiWordNet. Only behaviors (level 2) are pre-assigned affective scores in the lexicon; for entities exhibiting these behaviors (level 1), affect is calculated on demand and in context.

3 Learning Behaviors from the Web

Veale & Hao make the simplifying but unjustified assumption that all stereotypical properties are adjectival in nature, and work from adjectival properties (as inventoried by WordNet) to the nouns that exemplify them by successively binding *ADJ* in the web query "as *ADJ* as a NOUN" to different adjectives. The resulting enfilade of queries is sent in rapid succession to the search engine Google. All bindings for NOUN are then automatically extracted from the results before being manually inspected. Here we instead use the *like* patterns "VERB+ing *like* a NOUN" and "VERB+ed *like* a NOUN", the preferred simile patterns to describe behavior.

Before performing a large-scale trawl of the web, we first conduct a pilot study on the Google n-grams (Brants & Franz, 2006), a database of contiguous n-word strings ($1 \leq n \geq 5$) with a web frequency of 40 or higher. The pattern "VERB+ing *like* a NOUN" matches over 8,000 4-grams, while "VERB+ed *like* a NOUN" matches almost 4,000. However, we find here a good deal of empty behaviors, such as *acting* (as in "*acting* like a baby" rather than "*acting* like an actor") and looking (as in "*looking* like a fool"). Indeed, just three empty behaviors – *acting*, *looking*/*looked* and *seemed* – account for almost 2,000 n-gram matches. Others, like *walking* and *eating*, are too general and merely allude to a stereotypical behavior (as in "*walking* like a penguin") rather than explicitly providing the specific behavior (*waddling*). Panning the n-gram matches yields a few hundred nuggets of stereotypical insight, such as "*circling* like a shark", "*salivating* like a dog" and "*clinging* like a leech". Our pilot study reveals that most instances of the *like*-simile patterns are not so specific and informative, making a large-scale web trawl with these patterns impracticable.

Instead we use a hypothesis-driven approach by first looking for attested mentions of a specific behavior with a given noun. Consider the target noun *zombie*: searching the Google 3-grams for matches to the patterns "DET VERB+ing *zombie*" and "DET VERB+ed *zombie*" yields the following hypotheses for the stereotypical behavior of zombies (numbers in parentheses are the frequencies of the corresponding 3-grams):

{ *decomposing*(1454), *devastating*(134), *shambling*(115), *rotting*(103), *ravaged*(98), *brainwashed*(94), *drooling*(84), *freaking*(83), *attacking*(80), *crazed*(79), *obsessed*(73), *infected*(72), *marauding*(71), *disturbed*(65), *wandering*(64), *reanimated*(54), *flying*(52), *flaming*(52), *revived*(47), *decaying*(41), *unexpected*(40)}

For each attested behavior we generate the corresponding *like*-simile, such as "*decomposing* like a zombie", and determine its frequency on the open web. Corresponding non-zero frequencies, obtained using Google, are as follows:

{ *drooling*(4480), *wandering*(3660), *shambling*(1240), *revived*(860), *rotting*(682), *brainwashed*(146), *reanimated*(141), *infected*(72), *flaming*(52), *decaying*(46), *decomposing*(8), *attacking*(7), *flying*(6), *freaking*(2), *obsessed*(3)}

Unlike Veale & Hao then, we do not use a relatively small (~ 2000) set of queries that are made wide-ranging through the use of wild-cards, but generate a very large set of specific queries (with no wild-cards) that each derive from an attested combination of a specific behavior and a specific noun. We are careful not to dispatch queries that contain empty behaviors, a list of which is determined during our initial pilot study with the Google n-grams. In all, we dispatch over 500,000 web queries, for the same number of attested combinations. No parsing of the web results is needed, and we need record only the total number of returned hits per query / combination.

4 Initial Evaluation

The 3-gram patterns "DET *VERB+ing* NOUN" and "DET *VERB+ed* NOUN" attest to the plausibility of a given noun-entity exhibiting a specific behavior, but they are only weakly suggestive about what is actually typical. As a basis for generating hypotheses about stereotypical behavior these patterns over-generate significantly, and less than 20% of our queries yield non-zero result sets when sent to the web.

As shown by the *zombie* example above, some web-attested behaviors are best judged as idiosyncratic rather than stereotypical. While *rotting*, *decaying* and *shambling* are just the kind of behaviors we expect of zombies, *freaking*, *flying* and *flaming* are ill-considered oddities that our behavior model can well do without. As one

might expect, such oddities tend to have lower web frequencies than more widely-accepted behaviors (like *drooling*), yet as noted in Kilgarriff (2007), raw web frequencies can be an unreliable guide to what is typical. Note for instance how *decomposing* has a low frequency of just 8 uses on the web (as indexed by Google).

Our web data exhibits another interesting phenomenon. Consider the noun-entities for which the behavior *brainwashed* is attested, both in the 3-grams ("a *brainwashed* NOUN") and on the web ("*brainwashed* like a NOUN"):

{ *cult*(1090), *zombie*(146), *robot*(9), *child*(7), *fool*(4), *kid*(4), *idiot*(3), *soldier*(2)}

Since cults often use brainwashing, we can consider *cult* to be stereotypical for this behavior. Zombies and robots, however, are not typically brainwashed, nor indeed are they even brainwash*able*. Rather, it is more accurate to suggest that the victims of brainwashing often resemble *robots* and *zombies*, and to the extent that brainwashing is made possible by being weak-minded, *fools*, *idiots*, *kids* and *children*. This appears to be an example of what Bolinger (1988) dubs *ataxis*, insofar that *brainwashed* is a "migrant modifier" that more aptly describes the target of the simile than the vehicle (*robot* or *zombie*). In this case we can sensibly conclude that *brainwashed* is a figurative behavior of *robots* and *zombies* (since they typically act like a brainwashed person) and is the kind of association we want in our behavioral model. In contrast, it would not be sensible to include *brainwashing* as part of the behavioral description of *fools*, *idiots*, *kids*, *children* or even *soldiers* (though the latter is perhaps debatable).

Ultimately, the stereotypicality of a behavioral association is a pragmatic *gut* issue for the designer of a lexico-semantic resource, one that cannot be automatically resolved by considering web frequency (or other statistical quantities) alone. As with the design of WordNet itself, it is best resolved by asking and answering the question "is this an association that I would want in my lexicon?". For this reason, we filter the results of the web harvesting process manually, to ensure that the final model contains only those behavioral descriptions that a human would consider typical. In the end then, our approach is a semi-automatic one: automated processes scour the Google n-grams for behavioral hypotheses, and seek supporting evidence for these hypotheses on the web (in the form of *like*-similes), before a manual pass is finally conducted to ensure the model has the hand-crafted quality of a resource like WordNet.

This semi-automation allows us to build a behavioral model of high quality and significant scale. The model maps 5649 unique nouns to 4256 unique behaviors and contains approx. unique 44,000 mappings overall. This behavior-based model is thus more than three times larger that the adjectival stereotype model reported in Veale & Hao (2007), which contains just over 12,000 noun-to-adjective mappings.

5 Next Steps

The behavioral model, which captures the stereotypical behavior of thousands of word-concepts from *apes* to *zombies*, can be viewed as a complementary addition not just to WordNet but to the other knowledge resources previously described. Most obviously it complements the adjectival-stereotype model of Veale & Hao, and integrating the two would yield a larger and richer resource, of stereotypical descriptions that combine both adjectival and behavioral properties. For example, in a combined model, the *baby* stereotype has the following 163 properties:

{ *delicate, squalling, weeping, baptized, adopted, startled, attentive, blessed, teeny, rocked, adorable, whining, bundled, toothless, placid, expected, rescued, treasured, new, sleepy, indulged, slumbering, weaned, supple, helpless, small, sleeping, animated, vulnerable, wailing, cradled, kicking, soft, rested, bellowing, blameless, grinning, screaming, tiny, cherished, reliant, thriving, loveable, guileless, mute, inexperienced, dribbling, unthreatening, nursed, angelic, bawling, beaming, naked, spoiled, scared, weak, squirming, blubbering, contented, smiling, wiggling, mewling, blubbing, sniffling, overtired, dimpled, loving, dear, tired, powerless, bewildered, peaceful, distressed, naive, wee, soiled, sucking, fussy, gurgling, vaccinated, heartwarming, pouting, constipated, drooling, quiet, wiggly, lovable, bare, weaning, suckling, cute, bald, whimpering, tender, pampered, incontinent, fleshy, charming, dependent, artless, fussing, flabby, babbling, warm, giddy, crawling, snoozing, hairless, cuddled, sweet, sobbing, squealing, wrapped, cooing, swaddled, laughing, toddling, fragile, innocent, moaning, gentle, terrified, precious, cranky, giggling, confused, cuddly, fat, ignorant, snoring, young, howling, screeching, shrieking, trusting, shivering, napping, resting, frightened, fresh, loved, demanding, chubby, adored, appealing, happy, tame, relaxed, wriggly, rocking, wriggling, conceived, clean, content, smooth, crying, submissive, bumbling, pink, sniveling, orphaned, harmless, pure*}

A cursory glance at this list reveals a rich description of the stereotypical baby, one that incorporates pleasant and unpleasant behaviors in ample numbers. It makes little sense to reduce such a nuanced description to a single measure of gross lexical affect, or to parcel the description into separate senses, each with its own subset of behaviors. Instead, the partitioning of the description can be done on demand, and in context, to suit the speaker's meaning: if a term is used pejoratively, we focus on those qualities that are typically unpleasant (*sniveling, submissive, cranky, whimpering*, etc.); if the term is used affectionately, we focus instead on those that typically convey affection (*blessed*, *delicate*, *pure*, *loved*, *trusting*, etc.); and so on. The affective rating of different qualities can be ascertained from any of the existing resources discussed earlier, with more or less success. Whissell's DoA is perhaps the most limited, while Mohammad & Turney's eight-dimensional model of emotion seems to possess the most nuance and power.

However, even basic properties and behaviors can be construed differently from one context to another. In some settings, for instance, *cunning* may be a positive description; in most others, it will likely be seen as negative. Many adjectival properties exhibit this duality of affect, such as *proud*, *tough*, *tame* and *fragile*, and the description of the stereotypical baby above contains many that could be used to compliment in one context and to insult in another.

For this reason, we shall concentrate next on the construction of a nuanced model of behavioral affect, in which the affective profile of a behavior or adjectival property (and thus of the entity that exhibits that behavior in context) changes in response to the intended meaning of the speaker. This model, which will form the second stage of the two-level affective lexicon outlined in the introduction, will allow us to see the positive in properties like *trusting*, *cunning* and *demanding*, and the negative in properties like *proud*, *unthreatening* and *innocent*, as the context demands.

The behavior model described here will be of considerable use in this goal, since we now have a reliable, large-scale means of determining which properties and behaviors co-occur with which. For instance, the baby stereotype tells us that *sniveling* co-occurs with *submissive* and *cranky* co-occurs with *whimpering*. From these co-occurrence patterns we have constructed a weighted graph of mutually-supporting behaviors and the entities that exhibit them. We are now conducting experiments on the use of PageRank and other graph-theoretic algorithms (as used in Rada & Tarau, 2004; Esuli & Sebastiani, 2007) to identify the most effective means of exploiting graph structure in the determination of affect.

6 Conclusions (for now)

The availability of large-scale lexical resources with rich sense inventories, like WordNet, has made it possible to move from affect lexica that assign gross affective properties directly to words (e.g., Whissell's DoA) to more sophisticated organizations that assign affect to particular word senses only (e.g., Wordnet-affect and SentiWordNet). This allows an affect lexicon to tease apart the aspects of a word/concept that carry positive or negative connotations (such as the indiscriminate and clumsy senses of *butcher*, or the heroic sense of *hero*, but not the sandwich sense of *hero*) and carefully assign the right properties to just the right senses.

But affect is not a phenomenon that respects sense boundaries, and the affective connotation of one sense of a word can easily spread to others. Thus, all senses of the word *butcher*, including the purely literal sense of a professional meat vendor, are tainted by the negative connotation of the metaphoric extension that describes an indiscriminate murderer. Likewise, the same sense of a word can be used with different affective connotations in different contexts, because even individual senses – what Cruse (1986:49) conceives of as "unitary 'quanta' of meaning" – denote complex objects with their own wide ranges of typical properties and expected behaviors. While sense distinctions allow us to make our affect lexica more precise, sense boundaries do not demarcate affect boundaries as surely as we would like. But the solution does not lie in sense proliferation, in which even more fine-grained senses are added to WordNet and other resources. Rather, it lies in an ability to dynamically construe new perspectives on existing senses as the context demands.

The work reported here is just one step in this direction. Only by adequately modeling what is typical and salient – that is, what is *stereotypical* – of the entities denoted by our words and their senses, can we begin to model how speakers in context subtly shift the boundaries of sense to effectively communicate an affective meaning.

7 Acknowledgements

This research was supported by the WCU (World Class University) program under the National Research Foundation of Korea, and funded by the Ministry of Education, Science and Technology of Korea (Project No: R31-30007).

References

Dwight Bolinger. (1988). Ataxis. In Rokko Linguistic Society (ed.), *Gendai no Gengo Kenkyu (Linguistics Today)*, Tokyo:1—17.

Thorsten Brants and Alex Franz. (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium.

D. A. Cruse. (1986). *Lexical Semantics*. London: Cambridge University Press.

Andrea Esuli and Fabrizio Sebastiani. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of LREC-2006, the 5th Conference on Language Resources and Evaluation*, 417-422.

Andrea Esuli and Fabrizio Sebastiani. (2007). PageRanking WordNet Synsets: An application to opinion mining. *Proceedings of ACL-2007, the 45th Annual Meeting of the Association for Computational Linguistics*.

Christiane Fellbaum (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Adam Kilgarriff. (2007). Googleology is Bad Science. *Computational Linguistics,* **33**(1):147-151.

Rada Mihalcea and Paul Tarau. (2004). TextRank: Bringing Order to Texts. *Proceedings of EMNLP-04, the 2004 Conference on Empirical Methods in Natural Language Processing*.

Saif M Mohammad and Peter D. Turney. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotional lexicon. *Proceedings of the NAACL-HLT 2010 workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Los Angeles, California.

Robert Plutchik. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research and experience*, **2**(1-2):1-135.

Peter D. Turney. (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". *Proceedings of ACL-2002, the 40th Annual Meeting of the Association for Computational Linguistics*. pp. 417–424.

Carlo Strapparava and Alessandro Valitutti. (2004). Wordnet-affect: an affective extension of Wordnet. *Proceedings of LREC-2004, the 4th International Conference on Language Resources and Evaluation*, Lisbon.

Tony Veale and Yanfen Hao. (2007). Making Lexical Ontologies Functional and Context-Sensitive. *Proceedings of ACL-2007, the 45th Annual Meeting of the Association of Computational Linguistics*, 57–64.

Cynthia Whissell. (1989). The dictionary of affect in language. In R. Plutchik and H. Kellerman (Eds.) *Emotion: Theory and research*. Harcourt Brace, 113-131.