A Context-sensitive, Multi-faceted model of Lexico-Conceptual Affect

Tony Veale
Web Science and Technology Division,
KAIST, Daejeon,
South Korea.
Tony.Veale@gmail.com

Abstract

Since we can 'spin' words and concepts to suit our affective needs, context is a major determinant of the perceived affect of a word or concept. We view this re-profiling as a selective emphasis or de-emphasis of the qualities that underpin our shared stereotype of a concept or a word meaning, and construct our model of the affective lexicon accordingly. We show how a large body of affective stereotypes can be acquired from the web, and also show how these are used to create and interpret affective metaphors.

1 Introduction

The builders of affective lexica face the vexing task of distilling the many and varied pragmatic uses of a word or concept into an overall semantic measure of affect. The task is greatly complicated by the fact that in each context of use, speakers may implicitly agree to focus on just a subset of the salient features of a concept, and it is these features that determine contextual affect. Naturally, disagreements arise when speakers do not implicitly arrive at such a consensus, as when people disagree about *hackers*: advocates often focus on qualities that emphasize curiosity or technical virtuosity, while opponents focus on qualities that emphasize criminality and a disregard for the law. In each case, it is the same concept, *Hacker*, that is being described, yet speakers can focus on different qualities to arrive at different affective stances.

Any gross measure of affect (such as e.g., that hackers are *good* or *bad*) must thus be grounded in a nuanced model of the stereotypical properties and behaviors of the underlying word-concept. As different stereotypical qualities are highlighted or de-emphasized in a given context – a particular metaphor, say, might describe *hackers as terrorists* or *hackers as artists* – we need to be able to re-calculate the perceived affect of the word-concept.

This paper presents such a stereotype-grounded model of the affective lexicon. After reviewing the relevant background in section 2, we present the basis of the model in section 3. Here we describe how a large body of feature-rich stereotypes is acquired from the web and from local n-grams. The model is evaluated in section 4. We conclude by showing the utility of the model to that most contextual of NLP phenomena – affective metaphor.

2 Related Work and Ideas

In its simplest form, an affect lexicon assigns an affective score – along one or more dimensions – to each word or sense. For instance, Whissell's (1989) *Dictionary of Affect* (or *DoA*) assigns a trio of 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*. Though Whissell's DoA is based on human ratings, Turney (2002) shows how affective valence can be derived from measures of word association in web texts.

Human intuitions are prized in matters of lexical affect. For reliable results on a large-scale, Mohammad & Turney (2010) and Mohammad & Yang (2011) thus used the *Mechanical Turk* to elicit human ratings of the emotional content of words. Ratings were sought along the eight dimensions identified in Plutchik (1980) as primary emotions: *trust , anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. Automated tests were used to exclude unsuitable raters. In all, 24,000+ word-sense pairs were annotated by five different raters.

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 75–79,
Jeju, Republic of Korea, 8-14 July 2012.

Liu *et al*. (2003) also present a multidimensional affective model that uses the six basic emotion categories of Ekman (1993) as its dimensions: *happy, sad, angry, fearful, disgusted* and *surprised*. These authors base estimates of affect on the contents of *Open Mind*, a common-sense knowledge-base (Singh, 2002) harvested from contributions of web volunteers. These contents are treated as sentential objects, and a range of NLP models is used to derive affective labels for the subset of contents (~10%) that appear to convey an emotional stance. These labels are then propagated to related concepts (e.g., *excitement* is propagated from *roller-coasters* to *amusement parks*) so that the implicit affect of many other concepts can be determined.

Strapparava and 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 and 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*. (*Wordnet-affect* also supports these three dimensions as a-labels, and adds a fourth, *ambiguous*). Esuli & Sebastiani (2007) improve on their affect scores by running a variant of the PageRank algorithm (see also Mihalcea and Tarau, 2004) on the graph structure that tacitly connects word-senses in WordNet to each other via the words used in their textual glosses.

These lexica 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 & Hao harvested tens of thousands of instances of this pattern from the Web, to extract sets of adjectival properties for thousands of commonplace nouns. They show that 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, then the estimated scores show a reliable correlation with the DoA's own scores. It thus makes computational sense to calculate the affect of a word-concept as a function of the affect of its most salient properties. Veale (2011) later built on this work to show how a property-rich stereotypical representation could be used for non-literal matching and retrieval of creative texts, such as metaphors and analogies.

Both Liu et al. (2003) and Veale & Hao (2010) argue for the importance of common-sense knowledge in the determination of affect. We incorporate ideas from both here, while choosing to build mainly on the latter, to construct a nuanced, two-level model of the affective lexicon.

3 An Affective Lexicon of Stereotypes

We construct the stereotype-based lexicon in two stages. For the first layer, a large collection of stereotypical descriptions is harvested from the web. As in Liu *et al*. (2003), our goal is to acquire a lightweight common-sense representation of many everyday concepts. For the second layer, we link these common-sense qualities in a *support graph* that captures how they mutually support each other in their co-description of a stereotypical idea. From this graph we can estimate pleasantness and unpleasantness valence scores for each property and behavior, and for the stereotypes that exhibit them.

Expanding on the approach in Veale (2011), we use two kinds of query for harvesting stereotypes from the web. The first, "as ADJ as a NOUN", acquires typical adjectival properties for noun concepts; the second, "VERB+*ing* like a NOUN" and "VERB+*ed* like a NOUN", acquires typical verb behaviors. Rather than use a wildcard * in both positions (ADJ and NOUN, or VERB and NOUN), which gives limited results with a search engine like Google, we generate fully instantiated similes from hypotheses generated via the Google n-grams (Brants & Franz, 2006). Thus, from the 3-gram "a drooling zombie" we generate the query "drooling *like* a zombie", and from the 3-gram "a mindless zombie" we generate "*as* mindless *as* a zombie".

Only those queries that retrieve one or more Web documents via the Google API indicate the most promising associations. This still gives us over 250,000 web-validated simile associations for our stereotypical model, and we filter these manually, to ensure that the lexicon is both reusable and

of the highest quality. We obtain rich descriptions for many stereotypical ideas, such as *Baby*, which is described via 163 typical properties and behaviors like *crying*, *drooling* and *guileless*. After this phase, the lexicon maps each of 9,479 stereotypes to a mix of 7,898 properties and behaviors.

We construct the second level of the lexicon by automatically linking these properties and behaviors to each other in a support graph. The intuition here is that properties which reinforce each other in a single description (e.g. "as *lush and green* as a jungle" or "as *hot and humid* as a sauna") are more likely to have a similar affect than properties which do not support each other. We first gather all Google 3-grams in which a pair of stereotypical properties or behaviors X and Y are linked via coordination, as in "*hot and humid*" or "*kicking and screaming*". A bidirectional link between X and Y is added to the support graph if one or more stereotypes in the lexicon contain both X and Y. If this is not so, we also ask whether both descriptors ever reinforce each other in Web similes, by posing the web query "*as X and Y as*". If this query has non-zero hits, we still add a link between X and Y.

Let **N** denote this support graph, and N(p) denote the set of neighboring terms to p, that is, the set of properties and behaviors that can mutually support a property p. Since every edge in **N** represents an affective context, we can estimate the likelihood that p is ever used in a positive or negative context if we know the positive or negative affect of enough members of N(p). So if we label enough vertices of **N** with + / – labels, we can interpolate a positive/negative affect for all vertices p in **N**.

We thus build a reference set **-R** of typically negative words, and a set **+R** of typically positive words. Given a few seed members of **-R** (such as *sad*, *evil*, etc.) and a few seed members of **+R** (such as *happy*, *wonderful*, etc.), we find many other candidates to add to **+R** and **-R** by considering neighbors of these seeds in **N**. After just three iterations, **+R** and **-R** contain ~2000 words each.

For a property p, we define $N^+(p)$ and $N^-(p)$ as

$$(1) \quad N^+(p) = N(p) \cap \mathbf{+R}$$

$$(2) \quad N^-(p) = N(p) \cap \mathbf{-R}$$

We assign pos/neg valence scores to each property p by interpolating from reference values to their neighbors in **N**. Unlike that of Takamura *et al*. (2005), the approach is non-iterative and involves no feedback between the nodes of **N,** and thus, no inter-dependence between adjacent affect scores:

$$(3) \quad \text{pos}(p) = \frac{|N^+(p)|}{|N^+(p) \cup N^-(p)|}$$

$$(4) \quad \text{neg}(p) = 1 - \text{pos}(p)$$

If a term S denotes a stereotypical idea and is described via a set of typical properties and behaviors *typical(S)* in the lexicon, then:

$$(5) \quad \text{pos}(S) = \frac{\sum_{p \in \text{typical}(S)} \text{pos}(p)}{|\text{typical}(S)|}$$

$$(6) \quad \text{neg}(S) = 1 - \text{pos}(S)$$

Thus, (5) and (6) calculate the mean affect of the properties and behaviors of S, as represented via *typical(S)*. We can now use (3) and (4) to separate *typical(S)* into those elements that are more negative than positive (putting an unpleasant spin on S in context) and those that are more positive than negative (putting a pleasant spin on S in context):

$$(7) \quad \text{posTypical}(S) = \{p \mid p \in \text{typical}(S) \wedge \text{pos}(p) > 0.5\}$$

$$(8) \quad \text{negTypical}(S) = \{p \mid p \in \text{typical}(S) \wedge \text{neg}(p) > 0.5\}$$

4 Empirical Evaluation

In the process of populating **+R** and **-R**, we identify a reference set of 478 positive stereotype nouns (such as *saint* and *hero*) and 677 negative stereotype nouns (such as *tyrant* and *monster*). We can use these reference stereotypes to test the effectiveness of (5) and (6), and thus, indirectly, of (3) and (4) and of the affective lexicon itself. Thus, we find that **96.7%** of the stereotypes in **+R** are correctly assigned a positivity score greater than 0.5 (*pos(S) > neg(S)*) by (5), while **96.2%** of the stereotypes in **-R** are correctly assigned a negativity score greater than 0.5 (*neg(S) > pos(S)*) by (6).

We can also use **+R** and **-R** as a gold standard for evaluating the separation of *typical*(S) into distinct positive and negative subsets *posTypical(S)* and *negTypical(S)* via (7) and (8). The lexicon contains 6,230 stereotypes with at least one property in **+R**∪**-R**. On average, **+R**∪**-R** contains 6.51 of the properties of each of these stereotypes, where, on average, 2.95 are in **+R** while 3.56 are in **-R**.

In a perfect separation, (7) should yield a positive subset that contains only those properties in

typical(*S*)∩**+R,** while (8) should yield a negative subset that contains only those in *typical*(*S*)∩**-R**.

Macro Averages **(6230 stereotypes)**	***Positive*** **properties**	***Negative*** **properties**
Precision	.962	.98
Recall	.975	.958
F-Score	.968	.968

Table 1. Average P/R/F1 scores for the affective retrieval of +/- properties from 6,230 stereotypes.

Viewing the problem as a retrieval task then, in which (7) and (8) are used to retrieve distinct positive and negative property sets for a stereotype *S*, we report the encouraging results of Table 1 above.

5 Re-shaping Affect in Figurative Contexts

The Google n-grams are a rich source of affective metaphors of the form *Target is Source*, such as "politicians are crooks", "Apple is a cult", "racism is a disease" and "Steve Jobs is a god". Let *src(T)* denote the set of stereotypes that are commonly used to describe T, where commonality is defined as the presence of the corresponding copula metaphor in the Google n-grams. Thus, for example:

src(racism) = {*problem, disease, poison, sin, crime, ideology, weapon, ...*}

src(Hitler) = {*monster, criminal, tyrant, idiot, madman, vegetarian, racist, ...*}

Let *srcTypical*(T) denote the aggregation of all properties ascribable to T via metaphors in *src*(T):

$$(9)\quad srcTypical(T) = \bigcup_{M \in src(T)} typical(M)$$

We can also use the *posTypical* and *negTypical* variants in (7) and (8) to focus only on metaphors that project positive or negative qualities onto T.

In effect, (9) provides a feature representation for a topic T as viewed through the prism of metaphor. This is useful when the source *S* in the metaphor *T is S* is not a known stereotype in the lexicon, as happens e.g. in *Apple is Scientology*. We can also estimate whether a given term *S* is more positive than negative by taking the average pos/neg valence of *src*(S). Such estimates are 87% correct when evaluated using **+R** and **-R** examples.

The properties and behaviors that are contextually relevant to the interpretation of *T is S* are given by

$$(10)\quad salient(T,S) = |srcTypical(T) \cup typical(T)| \cap |srcTypical(S) \cup typical(S)|$$

In the context of *T is S*, the figurative perspective $M \in src(S) \cup src(T) \cup \{S\}$ is deemed apt for T if:

$$(11)\quad apt(M, T,S) = |salient(T,S) \cap typical(M)| > 0$$

and the degree to which M is apt for T is given by:

$$(12)\quad aptness(M,T,S) = \frac{|salient(T, S) \cap typical(M)|}{|typical(M)|}$$

We can construct an interpretation for *T is S* by considering not just {S}, but the stereotypes in *src*(T) that are apt for T in the context of *T is S*, as well as the stereotypes that are commonly used to describe S – that is, *src*(S) – that are also apt for T:

$$(13)\quad interpretation(T, S) = \{M | M \in src(T) \cup src(S) \cup \{S\} \wedge apt(M, T, S)\}$$

The elements $\{M_i\}$ of *interpretation*(T, S) can now be sorted by $aptness(M_i\ T, S)$ to produce a ranked list of interpretations (M_1, M_2 … M_n). For any interpretation M, the salient features of M are thus:

$$(14)\quad salient(M, T,S) = typical(M) \cap salient(T,S)$$

So *interpretation*(T, S) is an expansion of the affective metaphor *T is S* that includes the common metaphors that are consistent with T *qua* S. For instance, "*Google is -Microsoft*" (where - indicates a negative spin) produces *{monopoly, threat, bully, giant, dinosaur, demon, ...}*. For each M_i in *interpretation*(T, S), *salient*(M_i, *T, S*) is an expansion of M_i that includes all of the qualities that are apt for T *qua* M_i (e.g. *threatening*, *sprawling*, *evil*, etc.).

6 Concluding Remarks

Metaphor is the perfect tool for influencing the perceived affect of words and concepts in context. The web application *Metaphor Magnet* provides a proof-of-concept demonstration of this re-shaping process at work, using the stereotype lexicon of §3, the selective highlighting of (7)–(8), and the model of metaphor in (9)–(14). It can be accessed at:

http://*boundinanutshell.com/metaphor-magnet*

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

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

Paul Ekman. 1993. Facial expression of emotion. American Psychologist, 48:384-392.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. Proc. 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. Proc. of ACL-2007, the 45th Annual Meeting of the Association for Computational Linguistics.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A Model of Textual Affect Sensing Using Real-World Knowledge. Proceedings of the 8th international conference on Intelligent user interfaces, pp. 125-132.

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 F. 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, CA.

Saif F. Mohammad and Tony Yang. 2011. Tracking sentiment in mail: how genders differ on emotional axes. Proceedings of the ACL 2011 WASSA workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Portland, Oregon.

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

Push Singh. 2002. The public acquisition of commonsense knowledge. Proceedings of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access. Palo Alto, CA.

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, Portugal.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientation of words using spin model. Proceedings of the 43rd Annual Meeting of the ACL, 133–140.

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

Veale, T. and Hao, Y. Making Lexical Ontologies Functional and Context-Sensitive. Proceedings of ACL-2007, the 45th Annual Meeting of the Association of Computational Linguistics, pp. 57–64. June 2007.

Veale, T. and Hao, Y. Detecting Ironic Intent in Creative Comparisons. Proceedings of ECAI'2010, the 19th European Conference on Artificial Intelligence, Lisbon. August 2010.

Veale, T. Creative Language Retrieval: A Robust Hybrid of Information Retrieval and Linguistic Creativity. Proceedings of ACL'2011, the 49th Annual Meeting of the Association of Computational Linguistics. June 2011.

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