Extracting Contextual Evaluativity

Kevin Reschke
University of California, Santa Cruz
kreschke@ucsc.edu

Pranav Anand
University of California, Santa Cruz
panand@ucsc.edu

Abstract

Recent work on evaluativity or sentiment in the language sciences has focused on the contributions that lexical items provide. In this paper, we discuss *contextual evaluativity*, stance that is inferred from lexical meaning and pragmatic environments. Focusing on assessor-grounding claims like *We liked him because he so clearly disliked Margaret Thatcher*, we build a corpus and construct a system employing compositional principles of evaluativity calculation to derive that *we* dislikes *Margaret Thatcher*. The resulting system has an F-score of 0.90 on our dataset, outperforming reasonable baselines, and indicating the viability of inferencing in the evaluative domain.

1 Contextual Evaluativity

A central aim of contemporary research on sentiment or evaluative language is the extraction of evaluative triples: ⟨evaluator, target, evaluation⟩. To date, both formal (e.g., Martin and White 2005, Potts 2005) and computational approaches (e.g., Pang and Lee 2008) have focused on how such triples are lexically encoded (e.g., the negative affect of *scoundrel* or *dislike*). While lexical properties are a key source of evaluative information, word-based considerations alone can miss pragmatic inferences resulting from context. (1), for example, communicates that the referent of *we* bears not only positive stance towards the referent of *him*, but also negative stance towards Margaret Thatcher:

(1) We liked him because he so clearly disliked Margaret Thatcher.
LEXICAL EVALUATIVITY: ⟨we, him, +⟩; ⟨he, M.T., -⟩
CONTEXTUAL EVALUATIVITY: ⟨we, M.T., -⟩

This paper argues for a compositional approach to contextual evaluativity similar to the compositional methods adopted for lexical evaluativity in Moilanen and Pulman (2007) and Nasukawa and Yi (2003). At the the heart of the approach is the treatment of verbal predicates (*dislike* in (1)) as evaluativity functors which relate argument/entity-level evaluativity to event-level evaluativity.

As discussed in §2, the utitlity of such a model surfaces in cases where the event-level evaluativity is known from context, and thus new information about the contextual evaluativity of the event participants (e.g. Margaret Thatcher) can be inferred. Consequently, the empirical focus of this paper is on structures like (1), where the second clause provides grounds for the sentiment encoded in the first, and hence has a predictable event-level evaluation from the first clause's evaluator. In §3 we describe the collection and annotation of a corpus of such assessment-grounding configurations from large-scale web data. This annotated corpus serves as a test bed for experimental evaluation of various implementations of the proposed compositional approach. The results of these experiments (§4) strongly support a compositional approach to contextual evaluativity inference. A simple compositional algorithm based on a small, manually created evaluativity functor lexicon demonstrated significantly better precision than non-compositional baselines. Moreover, a method for automatically expanding coverage to novel predicates based on similarity with the manually created lexicon is shown to increase recall dramatically with modest reduction in precision.

2 A Framework For Inferring Contextual Polarity

Evaluativity is concerned with determining private states (e.g., judgment or emotion) that a particular evaluator bears towards a target entity, event, or proposition. This may be represented as a three place

Table 1: Evaluativity functors for verbs of having, withholding, disliking, and liking

x	y	E_{have}	E_{lack}	E_{wthhld}	E_{dprv}	E_{spr}	$E_{dislike}$	E_{like}
+	+	+	-	-	-	#	-	+
+	-	-	+	+	#	+	+	-
-	+	-	+	+	+	#	-	+
-	-	+	-	-	#	-	+	-
x have/lack *y*		*a* withhold/deprive/spare *x* of *y*					*x* dislike/like *y*	

relation, $R \subseteq D_e \times D_\alpha \times D_\mathcal{E}$, where α is of variable type and $\mathcal{E}$ is the type of evaluative stance, assumed here to be binary. Lexical approaches to evaluativity (see Pang and Lee 2008 for a review) have focused on those relations that are determinable from word-internal meaning alone. For example, describing an event e as *coddling* gives rise to two triples: $\langle \text{AGENT}(e), \text{PATIENT}(e), + \rangle$ and $\langle \text{SPEAKER}, e, - \rangle$.[1] These lexical inferences then become part of the feature set for classifying phrasal stance (e.g., the author's overall evaluativity in a sentence). A contrasting line of research (Moilanen and Pulman 2007, Nasukawa and Yi 2003) analyzes phrasal stance as a compositional product of the polarities toward event participants. For example, the evaluative polarity of the speaker toward the event in (2a) is positively correlated with the polarity toward the subject, and negatively so in (2b).

(2) a. My {ally, enemy} was deprived shelter.
b. My {ally, enemy} was spared a dangerous mission.

Compositional proposals rely on mapping each n-ary predicate P an n-ary evaluativity functor $E_P : D_\mathcal{E}{}^n \rightarrow D_\mathcal{E}$. Anand and Reschke (2011) argue that evaluativity functors largely group into classes, depending on whether the predicates in question entail final states of possession and/or affectedness. For example, the functors for predicates of withholding, including *deprive* and *spare*, are partial cases of the functor for *lack* (partiality reflects lexical idiosyncracies about e.g., deprivation and positive objects), as shown in Table 1.

While compositional systems are designed to compute phrasal stances bottom-up, their calculi straightforwardly allow inference to participant polarities as well, assuming knowledge of the event polarity and all but one participant. Consider the sentence *He disliked Margaret Thatcher*. By the evaluativity conditions in Table 1, $E_{dislike}$ is positive iff the evaluator has negative evaluation of Thatcher. Thus, given knowledge of the event polarity, we can infer the evaluator's stance with respect to Thatcher. In (1), this information is provided by the preceding assessing clause (+, from E_{like}). As the second clause serves as grounds for the assessment in the first clause, the event described in the second clause is predictably also assessed as + by the evaluator *we*. In our experiments we exploited this construction in particular, but the general procedure does not require it (thus, for example, evaluative adverbs such as *fortunately* and *regrettably* could provide an additional construction type). This procedure is sketched for (1) below:

(3) We liked$_{e_{like}}$ him because he so clearly disliked$_{e_{dislike}}$ Margaret Thatcher.
LEXICAL EVALUATIVITY: ⟨we, him, +⟩; ⟨he, M.T., -⟩
PRAGMATIC INFERENCE: ⟨we, e$_{dislike}$, +⟩ ($e_{dislike}$ justifies ⟨we, him, +⟩)
COMPOSITIONAL INFERENCE: $E_{dislike}(+, y) = +$ iff $y = +$
therefore, y is regarded as $+$, or ⟨we, M.T., -⟩

Note that for this application, we may simplify the compositional picture and treat functors as either preservers or reversers of the polarity of the object of interest, as is done in Moilanen and Pulman (2007) and Nasukawa and Yi (2003): preservers (such as verbs of liking) match the object polarity with the event polarity, and reverses negate it.

When the assessing clause evaluator is not affiliated with the speaker, this procedure can produce markedly different results from lexical markers (which often show speaker evaluativity). Thus, in (4), the speaker's assessment of Obama's cuts (indicated by the lexical *much-needed*) stands in sharp contrast with NASA's (determined by inference):

[1] Here, we simplify regarding potential evaluators outside of the speaker.

(4) NASA got angry at Obama because he imposed some much-needed cuts.
LEXICAL EVALUATIVITY: ⟨NASA, Obama, -⟩; ⟨SPEAKER, some much needed cuts, +⟩
CONTEXTUAL EVALUATIVITY: ⟨NASA, some much needed cuts, -⟩

The assessment-grounding configuration in (1) and (4) is highly productive. Behaviorally, *implicit causality* predicates (including predicates of assessment, as well as praise and scolding) are frequently understood by experimental subjects as describing an event involving the assessment target, especially when followed by *because* (Garvey and Carmazza, 1974; Koornneef and van Berkum, 2006). In addition, Somasundaran and Weibe (2009) exploited a similar construction to gather reasons for people's product assessments from online reviews. These together suggest that such constructions could be simultaneously high-precision sources for evaluativity inference and easily obtainable from large corpora.

3 Data Gathering and Annotation

We developed a corpus of assessment-grounding excerpts from documents across the web to evaluate the potential of the framework in §2. 73 positive and 120 negative assessment predicates (*like*, *adore*, *hate*, *loathe*, etc.) were selected from the MPQA subjectivity lexicon (Wilson et al., 2005). These were expanded accross inflectional variants to produce 826 assessment templates, half with explicit *because*, half without (e.g. *terrified by* X *because he*). These templates were filled with personal pronouns and the names of 26 prominent political figures and issued as websearch queries to the Yahoo! Search API.[2] A total of 440,000 webdocument results were downloaded and processed using an 1152 core Sun Microsystems blade cluster. The relevant sentences from each document were extracted, and those under 80 characters in length were parsed using the Stanford Dependency Parser.[3]

This produced 60,000 parsed assessment-grounding sentences, 6,000 of which (excluding duplicates) passed the additional criterion that the grounding clause should contain a verb with a direct object. This restriction ensured that each item in our corpus had a target for contextual polarity inference. An additional 3,300 cases were excluded because the target in the grounding clause shared possible coreference with the experiencer (subject) of the assessment clause. We avoided these coreferring cases because, from the perspective of a potential application, inferences about an experiencer's stance towards himself are less valuable than inferences about his stance towards others. Finally, the list was manually shortened to include only those sentences marked as assessment-grounding configurations according to two annotators ($\kappa = 0.82$); the classification task of whether this pragmatic connection occurs is beyond the scope of this paper. 57% of the data was removed in this pass, 14% from tokens with *because* and 43% from tokens without. Implicit causality verbs not followed by *because* have been shown experimentally to give rise to a much weaker preference for justification (Au, 1986), and this is confirmed in our corpus search. The result of this procedure was a final corpus size of 1,160.

The corpus was annotated for inferred contextual polarity. One of the authors and another annotator coded sentences for evaluator stance toward the object (+,-, unknown); agreement was high: $\kappa = 0.90$. The 48 unresolved cases were adjudicated by a third annotator. 27 cases were uniformly judged unknown, involving predicates of change, disclosure (*reveal, expose*), and understanding (*know*). These were removed from the corpus, leaving 1,133 sentences for training and testing.

4 System and Experimental Results

Restricting ourselves to the assessment-grounding configuration discussed above, we treat contextual polarity inference as a binary classification problem with two inputs: the INPUT EVENT event-level polarity (derived from the assessment clause) and the main verb of the grounding clause (henceforth FUNCTOR VERB). The goal of the classifier is to correctly predict the polarity of the target NP (direct object to the functor verb) given these inputs.

[2] http://developer.yahoo.com/search/

[3] http://nlp.stanford.edu/software/lex-parser.shtml

Table 2: Examples of verbs marked as preserver/reverser and their sources

EXAMPLE	CLASS	SOURCE
reward	preserver	MPQA subj. lex.
hamper	reverser	MPQA subj. lex.
tutor	preserver (benefit)	FrameNet
batter	reverser (injury)	FrameNet

Table 3: Performance of systems and baselines for contextual evaluativity classification

SYSTEM	PREC.	RECALL	F-SCORE
B-Functor	0.39	0.24	0.30
B-Input	0.69	1.0	0.82
B-MLE	0.75	1.0	0.86
SYS	0.88	0.57	0.69
SYS-MPQA	0.88	0.24	0.38
SYS-Frame	0.89	0.41	0.56
SYS+Maj	0.82	1.0	**0.90**
SYS+Sim	0.84	0.97	**0.90**

As mentioned in §2, we may categorize the functor verbs in our lexicon into preservers and reversers. Two sources populate our lexicon. First, positively subjective verbs from the MPQA subjectivity lexicon were marked as preservers and negatively subjective verbs were marked as reversers (1249 verbs total). For example, $E_{dislike}$ is a reverser. Second, 487 verbs were culled from FrameNet (Ruppenhofer et al., 2005) based on their membership in six entailment classes: verbs of injury, destruction, lacking, benefit, creation, and having. Class membership was determined by identifying 124 FrameNet frames aligning with one or more classes, then manually selecting from these frames verbs whose class membership was unambiguous. Verbs of benefit, creation, and having were marked as preservers. Verbs of injury, destruction, and lacking were marked as reversers (Table 2). Our system (SYS) classifies objects in context as follows: If the functor verb is a preserver, the target NP is assigned the same polarity as the input event polarity. If the functor verb is a reverser, the target NP is assigned the opposite of the input event polarity. This procedure is modulated by the presence of negation, as detected by a *neg* relation in the dependency parse. Under negation, a preserver acts like a reverser, and vice versa.

We tested the performance of this system (SYS) on our annotated corpus against three baselines. The first baseline (B-Functor) attempts to determine the importance of the input event to the calculation. It thus ignores the preceding context, and attempts to classify the target object from the functor verb directly, based on the verb's polarity in the MPQA subjectivity lexicon. It has poor precision and recall,[4] reflecting both the importance of the assessment context for object polarity and the fact that the functor verbs are often not lexically sentiment bearing (e.g., predicates of possession). The second baseline (B-Input), conversely, ignores the functor verb and uses the input event polarity as listed in the MPQA lexicon (modulo negation) for object classification. The purpose of this baseline is to approximate a classifier that predicts target polarity solely from the global/contextual polarity of the preceding clause. This has sharply increased precision, indicating contextual information's importance. The third baseline (B-MLE) picked the majority object class (+), and had the highest precision, indicating the general bias in our corpus for positive objects. Table 3 shows the performance (precision vs. recall) of our system compared to the three baselines. Its precision is significantly higher, but its F-score is limited by the lower coverage of our manually constructed lexicon. SYS-MPQA and SYS-Frame show the performance of the system when the functor lexicon is limited to the MPQA and Framenet predicates, respectively. Both are high precision sources of functor prediction, and pick out somewhat distinct predicates (given the recall gain of combining them). SYS+Maj and SYS+Sim are attempts to handle the low recall of SYS caused by functor verbs in the test data which aren't in the system's lexicon. SYS+Maj simply assigns these out-of-vocabulary verbs to the majority class: preservers. SYS+Sim classifies out-of-vocabulary verbs as preservers or reversers based on their relative similarity to the known preservers and reversers selected from FrameNet – an unknown verb is categorized as a preserver if its average similarity to preservers is greater than its average similarity to reversers. Similarity was determined according to the Jiang-Conrath distance measure (Jiang and Conrath, 1997), which based on links in WordNet (Fellbaum, 1998). (Note: this process cannot occur for words not found in WordNet – e.g. misspellings – hence the

[4]Low recall occurs when items are left unclassified due to out-of-vocabulary functor verbs. Low precision occurs when a + item is classified as – or vice versa.

less than perfect recall). These two systems outperform all baselines, but have indistinguishable F-scores (if misspellings are excluded, SYS+Sim has a Recall of 0.99 and F-score of 0.91).

Most of the precision errors incurred by our systems were syntactic: incorrect parsing, incorrect extraction of the object, or faulty negation handling (e.g., negative quantifiers or verbs). 26% of errors are due to word-sense disambiguation. The verbs *spoil* and *own* each have positive and negative uses (*own* can mean *defeat*), but only one sense was registered in the lexicon, leading to errors. The lion's share of these errors (22%) were due to the use of *hate* and similar expressions to convey jealousy (e.g. *I was mad at him because he had both Boardwalk and Park Place*). In these scenarios, although the assessment is negative, the event-level polarity of the grounding clause event type is positive (because it is desired), a fact which our current system cannot handle. One way forward would be to apply WSD techniques to distinguish jealous from non-jealous uses of predicates of dislike.

5 Conclusion

We have described a system for the extraction of what we termed contextual evaluativity – evaluations of objects that arise from the understanding of pragmatic inferences. This system, once we incorporate procedures to automatically infer evaluativity functor class, significantly outperforms reasonable baselines on a corpus of assessor-grounding extracts from web documents. The system operates by running a compositional approach to phrasal evaluativity in reverse, and is thus an instance of the potential computational value of such treatments of evaluativity.

References

Anand, P. and K. Reschke (2011). Verb classes as evaluativity functor classes. In *Interdisciplinary Workshop on Verbs*. to appear.

Au, T. K. (1986). A verb is worth a thousand words. *Journal of Memory and Language 25*, 104–122.

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

Garvey, C. and A. Carmazza (1974). Implict causality in verbs. *Linguistic Inquiry 5*, 459–484.

Jiang, J. and C. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*.

Koornneef, A. W. and J. J. A. van Berkum (2006). On the use of verb-based implicit causality in sentence comprehension. *Journal of Memory and Language 54*, 445–465.

Martin, J. R. and P. R. R. White (2005). *Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Moilanen, K. and S. Pulman (2007). Sentiment composition. In *Proceedings of RANLP 2007*.

Nasukawa, T. and J. Yi (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*.

Pang, B. and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*(1-2), 1–135.

Potts, C. (2005). *The Logic of Conventional Implicature*. Oxford University Press.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, and C. R. Johnson (2005). Framenet ii: Extended theory and practice. Technical report, ICSI Technical Report.

Somasundaran, S. and J. Weibe (2009). Recognizing stances in online debates. In *Proceedings of ACL-47*, pp. 226–234.

Wilson, T., J. Weibe, and P. Hoffman (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of EMNLP-05*.