POLITICAL-ADS: An annotated corpus of event-level evaluativity

Kevin Reschke
Department of Computer Science
Stanford University
Palo Alto, CA 94305 USA
reschkek@gmail.com

Pranav Anand
Department of Linguistics
University of California, Santa Cruz
Santa Cruz, CA 95064 USA
panand@ucsc.edu

Abstract

This paper presents a corpus targeting evaluative meaning as it pertains to descriptions of events. The corpus, POLITICAL-ADS is drawn from 141 television ads from the 2008 U.S. presidential race and contains 3945 NPs and 1549 VPs annotated for scalar sentiment from three different perspectives: the narrator, the annotator, and general society. We show that annotators can distinguish these perspectives reliably and that correlation between the annotator's own perspective and that of a generic individual is higher than those with the narrator. Finally, as a sample application, we demonstrate that a simple compositional model built off of lexical resources outperforms a lexical baseline.

1 Introduction

In the past decade, the semantics of evaluative language has received renewed attention in both formal and computational linguistics (Martin and White, 2005; Potts, 2005; Pang and Lee, 2008; Jackendoff, 2007). This work has focused on evaluativity at either the lexical level or the phrasal/event level stance, without bridging between the two. A parallel tradition of compositional event polarity ((Nasukawa and Yi, 2003; Moilanen and Pulman, 2007; Choi and Cardie, 2008; Neviarouskaya et al., 2010)) has grown up analogous to approaches to compositionality in formal semantics: event predicates are not of constant polarity, but provide functions from the polarities of their arguments to event polarities. Little work exists assessing the relative advantages of a compositional account, in part because no resource annotating both NP level polarity and event-level polarity in context exists. This paper introduces such a corpus, POLITICAL-ADS, a collection of 2008 U.S. presidential race television ads with scalar sentiment annotations at the NP and VP level. After describing the corpus creation and characteristics in sections 3 and 4, in section 5, we show that a compositional system achieves an accuracy of 84.2%, above a lexical baseline of 65.1%.

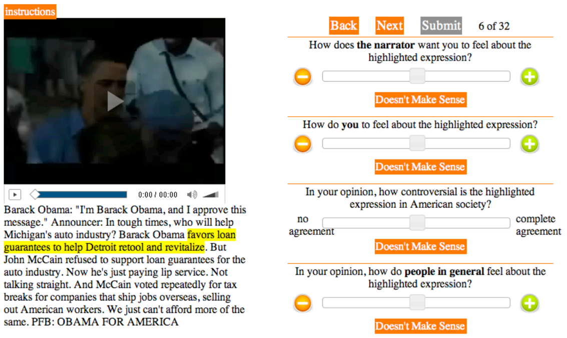
2 Background

While many sentiment models handle negation quasi-compositionally (Pang and Lee, 2008; Polanyi and Zaenen, 2005), Nasukawa & Yi (Nasukawa and Yi, 2003) first noted that predicates like *prevent* are "flippers", conveying that their subject and object have opposite polarity – since *trouble* is negative, something that *prevents trouble* is good. Recent work has expanded that idea into a fully compositional system (Moilanen and Pulman, 2007; Neviarouskaya et al., 2010). Moilanen and Pulman construct a system of compositional rules that builds polarityin terms of a hand-built lexicon of predicates as flippers or preservers. However, this system conflates two different assessment perspectives, that of the Narrator and of some mentioned NP (NP-to-NP perspective). The latter include psychological predicates such as *love* and *hate*, and those of admiration or censure (e.g., *admonish*, *praise*). Thus, they would mark *John dislikes scary movies* as negative, a correct NP-to-NP claim, but not necessarily correct for the Narrator. Recognizing this, Neviarouskaya et al. (Neviarouskaya et al., 2010) develop a pair of

Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pages 84–88,
Jeju, Republic of Korea, 12 July 2012. ©2012 Association for Computational Linguistics

Announcer: In tough times, who will help Michigan's auto industry? Barack Obama favors loan guarantees to help Detroit retool and revitalize. But John McCain refused to support loan guarantees for the auto industry. Now he's just paying lip service. Not talking straight. And McCain voted repeatedly for tax breaks for companies that ship jobs overseas, selling out American annotators. We just can't afford more of the same.

Figure 1: Transcript of POLITICAL-ADS ad #57



Barack Obama: "I'm Barack Obama, and I approve this message." Announcer: In tough times, who will help Michigan's auto industry? Barack Obama favors loan guarantees to help Detroit retool and revitalize. But John McCain refused to support loan guarantees for the auto industry. Now he's just paying lip service. Not talking straight. And McCain voted repeatedly for tax breaks for companies that ship jobs overseas, selling out American workers. We just can't afford more of the same. PFB: OBAMA FOR AMERICA

How does **the narrator** want you to feel about the highlighted expression?

How do **you** to feel about the highlighted expression?

In your opinion, how controversial is the highlighted expression in American society?

In your opinion, how do **people in general** feel about the highlighted expression?

Figure 2: POLITICAL-ADS annotation interface

compositional rules over both perspectives. Importantly, neither of these approaches have been validated against a sufficiently nuanced dataset. Mailanen and Pulman test against the SemEval-07 Headlines Corpus, which asks annotators to give an overall impression of sentiment. This approach allows a headline such as *Outcry in N Korea 'nuclear test'* to be marked negative, even though outcry over military provocations is arguably good. Similarly, Neviarouskaya et al. evaluate only against NP-to-NP data as well. While the MPQA corpus (Wiebe et al., 2005), which annotates the source of each sentiment annotation, separates these two sentiment sources, work trained on it has not (Choi and Cardie, 2008; Moilanen et al., 2010). In addition, existing annotation schemes are not designed to tease apart perspectival differences. For example, MPQA includes a notion of Narrator-oriented evaluativity, but it does not include the perspectives of you and the general public.

3 The corpus

POLITICAL-ADS, is drawn from politics, a rich and recently evolving domain for evaluativity research that we hypothesized would involve a high volume of sentiment claims subject to perspectival differences. POLITICAL-ADS is a collection of 141 television ads that ran during the 2008 U.S. presidential race between Democratic candidate Barack Obama and Republican candidate John McCain. The collection consists of 81 ads from Democratic side and 60 ads from Republican side. Figure 1 provides a sample transcript.

Each transcript was parsed using the Stanford Parser and all NPs and VPs excluding those headed by auxiliaries were extracted. VP annotations were assumed to represent phrasal/event-level polarity and NP ones argument-level polarity. The annotation interface is shown in Figure 2. Annotators were shown a transcript and a movie clip, and navigated through the NPs and VPs within the document. At each point they were asked to rate their response on a [-1,1] scale for the following four questions about the highlighted expression: 1) how the narrator wants them to feel; 2) how they feel; 3) how people in general feel; 4) how controversial the issue is (included to test the whether sense of controversy yields sharper differences between the various assessment perspectives). Finally, because phrases were not prefiltered, a 'Doesn't Make Sense' button was provided for each question.

206 annotators on Mechanical Turk completed 985 transcripts at $0.40 per transcript; each transcript was annotated by an average of 4.8 different annotators living in the U.S. We then filtered annotators by 200 phrases we deemed relatively uncontroversial in 20 randomly selected transcripts. To do this, we scored each annotator in terms of the absolute difference between their mean response and the median (each annotator's scores were first normalized by mean absolute value) in the Narrator question. We found when we thresholded annotators at a score above 0.5, agreement with our gold standard was 83.5% and dropped substantially afterwards. This threshold excluded 74 annotators, leaving 132 high-quality, or HQ, annotators (the full data is available in the corpus).

The corpus consists of 5494 phrases (1549 VPs and 3945 NPs) annotated 6.3 times on average, for a total of $34,692$ annotations (9800 VP and 24892 NP). Each phrase was annotated by at least 3 HQ annotators (average 3.9 annotators), and such annotators contributed 5960 VP and 15238 NP annotations. Of these, 12.1% HQ NP and 5.4% of HQ VP responses were marked as 'Doesn't Make Sense' (DMS) for the narrator question. In general, controversy and narrator questions had the highest and lowest rates of DMS, respectively; NPs showed higher response rates than VPs; and HQ annotators had higher rates of button presses.[1] In sections 4 and 5, we will ignore the DMS responses.

4 Corpus Findings

Table 1 provides summary statistics for the corpus. Across the board, the three perspective questions averaged close to 0, and in general HQ annotators are closer to 0 (non-HQ annotators tended to provide positive responses). VPs had slightly higher variance than NPs, at marginal probability ($p < .04$), suggesting that VP responses were more extreme than NP ones. You and Generic assessments are highly correlated (Pearson's $\rho = 0.85$), but Narrator is less so ($\rho = .76/.74$). All three are weakly correlated with Controversy ($\rho = .25/.26/.29$ for Narr., You, Gen., respectively). Narrator has the highest standard deviations for the raw data, but the lowest for the normed data. In the raw data, many annotators recognized the narrators intensely partisan views and rated accordingly ($|x| > 0.8$), but were more tempered when providing their perspective ($|x| \sim 0.35$), leading to lower σ. This intensity difference is factored out in normalization, yielding the opposite pattern.

The response data was collected from our annotators in scalar form, but applications (e.g., evaluative polarity classification) it is the polarity of the response that matters. Ignoring magnitude, Table 3 shows the polarity breakdown for all HQ phrasal annotations. Positive responses are the dominant class across the board. Neutral responses are less frequent for Narrator than for the other types. NPs have fewer negatives and more neutrals than VPs.

Table 2 shows average standard deviations (i.e., agreement) by worker, question, and XP type. Note both that NPs show less variance than VPs and that non-HQ annotators less than HQ annotators (non-HQ annotators gave more 0 responses).

COND	ALL	HQ ONLY	
	RAW	RAW	NORMED
Narr.	.10 (.45)	.05 (.62)	.08 (.87)
You	.10 (.34)	.06 (.46)	.09 (.85)
Gen.	.10 (.33)	.05 (.45)	.08 (.86)
Contr.	.17 (.22)	.13 (.30)	.17 (.60)

Table 1: Mean response by category and worker type

COND	HQ ANNOTATORS					
	RAW			NORMED		
	ALL	VP	NP	ALL	VP	NP
Narr.	.69	.75	.67	.96	1.06	.93
You	.57	.63	.55	.99	1.12	.94
Gen.	.53	.58	.51	.99	1.13	.94
Contr.	.53	.58	.51	1.01	1.15	.96
	ALL ANNOTATORS					
	ALL	VP	NP			
Narr.	.63	.68	.62			
You	.54	.59	.53			
Gen.	.52	.56	.51			
Contr.	.54	.56				

Table 2: Average Standard Deviations For HQ and all annotators

5 Comparing lexical and compositional treatments

While compositional models of event-level evaluativity are logically defensible, the extent to which these models apply in the wild is an open question. Because other compositional lexicons are not freely available, we used the system described in (Reschke and Anand, 2011), which induces flippers and preservers from the MPQA subjectivity lexicon and FrameNet (Ruppenhofer et al., 2005). The MPQA lexicon is a collection of over 8,000 words marked for polarity. Our functor lexicon uses the following heuristic: verbs marked positive in MPQA are preservers; verbs marked negative are flippers. For example, *dislike* has negative MPQA polarity; therefore, it is marked as a flipper in our lexicon. This gives us 1249 predicates: 869 flippers and 380 preservers. 329 additional verbs were added from FrameNet according to their membership in five en-

[1] In a QUESTION + PHRASE TYPE + QUESTION + ANNOTATOR TYPE linear model with annotator as a random effect, all of the above effects are significant. This was the simplest model according to χ^w model comparison.

COND	POL	VP	NP
Narr.	+	2874 (51%)	6877 (51%)
	-	2654 (47%)	5590 (42%)
	0	111 (2%)	932 (7%)
You	+	2714 (49%)	6573 (50%)
	-	2466 (45%)	4967 (38%)
	0	337 (6%)	1575 (12%)
Gen.	+	2615 (48%)	6350 (49%)
	-	2541 (48%)	5125 (39%)
	0	332 (6%)	1558 (12%)
Contr.	+	3095 (57%)	6522 (51%)
	-	1755 (32%)	4159 (33%)
	0	558 (10%)	2051 (16%)

Table 3: Polarity breakdowns for HQ annotations

tailment classes (Reschke and Anand, 2011): verbs of injury/destruction, lacking, benefit, creation, and having. 124 frames across these classes were identified, and then verbs of benefit, creation, and having (*aid, generate, have*) were marked as preservers and the complement set (*forget, arrest, lack*) as flippers. As a lexical baseline, the MPQA polarity of each verb was used – flippers correspond to baseline negative events and preservers to positive ones.

A 635 VP test subset of POLITICAL-ADS was constructed by omitting intransitive VPs and VPs with non-NP complements. Gold standard labels were determined from average normed HQ annotator data. This yielded 329 positive, 284 negative, and 2 neutral events. NPs, determined similarly, divided into 393 positive, 230 negative, and 12 neutral. Of the 635 VPs in the test set, only 272 (43.5%) are in our FrameNet/MPQA lexicon and we hence compare the two systems on this subset. On this subset, the compositional system has an accuracy of 84.2%, while the lexical baseline has an accuracy of 65.1%; there were 72 instances where the compositional model outperformed the lexical baseline and 22 where the lexical outperformed the compositional. Typical examples where the compositional system won involve MPQA negatives like *break*, *cut*, and *hate* and positives like *want* and *trust*. The lexical model marks VPs like *breaks the grip of foreign oil* and *want a massive government* as negative and positive, respectively – because the NPs in question are negative, the answers should be reversed. In contrast, the lexical model wins on cases like *grow the economy* and *reform Wall Street* correct. These exemplify a robust pattern in the errors: cases where the event is marked positive while the NP is marked negative. In examples like *grow Washington*, the idea that *grow* is a preserver is reasonable. However, in *grow the economy*, the negativity of the economy is arguably measuring the state of some constant entity. While *reform* is marked positive in MPQA, it is arguably a reverser; this shows the problems with our lexicon induction.

At an intuitive level, we expect agent evaluativity to mirror event-level evaluativity because positive/negative entities tend to commit positive/negative acts, and this is borne out. For flippers or preservers, the average VP evaluativity is correlated with the average subject evaluativity. For flippers the correlation is 0.57; for preservers it is 0.52. Although our model ignored subject evaluativity, we performed a generalized linear regression with subject and object evaluativity as predictors and event-level evaluativity as outcome. For flippers the regression coefficients were 0.52 for subject ($p < 4e-4$) and -0.52 for object ($p < 1e-5$). For preservers the coefficients were 0.27 ($p < 1e-5$) for subject and 0.93 for object ($p < 2e-7$). Thus, subject polarity is an important factor for flipper events (e.g., *the hero/villain defeated the enemy*, but less so for preservers (e.g. *the hero/villain helped the enemy.*).

6 Conclusion

In this paper we have presented POLITICAL-ADS, a new resource for investigating the relationships between NP sentiment and VP sentiment systematically. We have demonstrated that annotators can reliably annotate political data with sentiment at the phrasal level from multiple perspectives. We have also shown that in the present data set that self-reporting and judging generic positions are highly correlated, while correlation with narrators is appreciably weaker, as narrators are seen as more extreme. We have also shown that the controversy of a phrase does not correlate with annotators' disagreements with the narrator. Finally, as a sample application, we demonstrated that a simple compositional model built off of lexical resources outperforms a purely lexical baseline.

References

Y. Choi and C Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of EMNLP 2008*.

Ray Jackendoff. 2007. *Language, consciousness, culture*. MIT Press.

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

Karo Moilanen and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of RANLP 2007*.

K. Moilanen, S. Pulman, and Y Zhang. 2010. Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression. In *Proceedings of WASSA 2010, EACI 2010*.

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

A. Neviarouskaya, H. Prendinger, , and M. Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of COLING 2010*.

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

L. Polanyi and A. Zaenen. 2005. Contextual valence shifters. in computing attitude and affect in text. In Janyce Wiebe James G. Shanahan, Yan Qu, editor, *Computing Attitude and Affect in Text: Theory and Application*. Springer Verlag, Dordrecht, The Netherlands.

Chris Potts. 2005. *The Logic of Conventional Implicature*. Oxford University Press.

K. Reschke and P. Anand. 2011. Extracting contextual evaluativity. In *Proceedings of ICWS 2011*.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. 2005. Framenet ii: Extended theory and practice. Technical report, ICSI Technical Report.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. In *Proceedings of LREC 2005*.