Identifying Subjective and Figurative Language in Online Dialogue

Stephanie M. Lukin, Luke Eisenberg, Thomas Corcoran, & Marilyn A. Walker
Natural Language and Dialogue Systems
Computer Science Department, SOE-3
University of California, Santa Cruz
slukin,leisenbe,tcorcora,mawalker@ucsc.edu

More and more of the information on the web is dialogic, from Facebook newsfeeds, to forum conversations, to comment threads on news articles. In contrast to traditional, monologic resources such as news, highly social dialogue is very frequent in social media, as illustrated in the snippets in Fig. 1 from the publicly available Internet Argument Corpus (**IAC**) (Walker et al., 2012). Utterances are frequently sarcastic, e.g., *Really? Well, when I have a kid, I'll be sure to just leave it in the woods, since it can apparently care for itself* (R2 in Fig. 1 as well as Q1 and R1), and are often nasty, (R2 in Fig. 1). Note also the frequent use of dialogue specific discourse cues, e.g. the use of *No* in R1, *Really? Well* in R2, and *okay, well* in Q3 in Fig. 1 (Fox Tree and Schrock, 1999; Bryant and Fox Tree, 2002; Fox Tree, 2010).

Quote **Q**, Response **R**	**Sarc**	**Nasty**
Q1: I jsut voted. sorry if some people actually have, you know, LIVES and don't sit around all day on debate forums to cater to some atheists posts that he thiks they should drop everything for. emoticon-rolleyes emoticon-rolleyes emoticon-rolleyes As to the rest of your post, well, from your attitude I can tell you are not Christian in the least. Therefore I am content in knowing where people that spew garbage like this will end up in the End. **R1**: No, let me guess . . . er . . . McDonalds. No, Disneyland. Am I getting closer?	1	-3.6
Q2: The key issue is that once children are born they are not physically dependent on a particular individual. **R2** Really? Well, when I have a kid, I'll be sure to just leave it in the woods, since it can apparently care for itself.	1	-1

Figure 1: Sample Quote/Response Pairs from 4forums.com with Mechanical Turk annotations for Sarcasm and Nasty/Nice. Highly negative values of Nasty/Nice indicate strong nastiness and sarcasm is indicated by values near 1.

We aim to automatically identify sarcastic and nasty utterances in unannotated online dialogue, extending a bootstrapping method previously applied to the classification of monologic subjective sentences by Riloff & Wiebe, henceforth R&W (Riloff and Wiebe, 2003; Thelen and Riloff, 2002). We look at both sarcastic and nasty dialogic turns as a way to explore generalization of the method. R&W's method creates a High-Precision, Cue-Based Classifier to be a first approximation on unannotated text. They improve their classifier by learning and bootstrapping patterns (Fig. 2).

We found that this bootstrapping method 'as is' is not

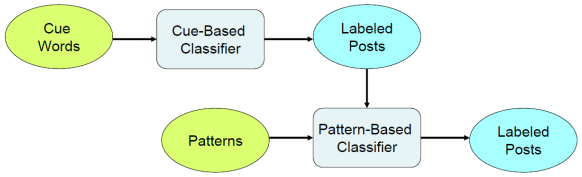

Figure 2: Bootstrapping Method

appropriate for our data because our Cue-Based Classifier yields a much lower precision than the bootstrapping requires. We have adapted the method to fit the sarcastic and nasty dialogic domain. Our method is as follows:

1. Explore methods for identifying sarcastic and nasty cue words and phrases in dialogues;
2. Use the learned cues to train a sarcastic (nasty) Cue-Based Classifier
3. Learn general syntactic extraction patterns from the sarcastic (nasty) utterances and define fine-tuned sarcastic patterns to create a Pattern-Based Classifier;
4. Combine both Cue-Based and fine-tuned Pattern-Based Classifiers to maximize precision at the expense of recall and test on unannotated utterances.

Cue Words. Sarcasm is known to be highly variable in form, and to depend, in some cases, on context for its interpretation (Sperber and Wilson, 1981; Gibbs, 2000; Bryant and Fox Tree, 2002). We elicit annotations from Mechanical Turk to identify sarcastic (nasty) cues in utterances from a development set. Turkers were presented with dialogic turns (a quote and its response) previously labeled sarcastic or nasty in the IAC by 7 different annotators, and were asked to identify sarcastic (nasty) or potentially sarcastic (nasty) phrases in the turn response. The Turkers then selected words or phrases from the response they believed could lead someone to believing the utterance was sarcastic or nasty. (Snow et al., 2008) measure the quality of Mechanical Turk annotations on common NLP tasks by comparing them to a gold standard. Pearson's correlation coefficient shows that very few Mechanical Turk annotators were required to beat the gold standard data, often less than 5. Because our sarcasm (nasty) task does not have gold standard data, we asked 100 annotators to participate in the pilot. For all unigrams, bigrams, and trigrams, interannotator agreement

SARC	PARAMS	P	R	F
Cue	θ_1 =2, θ_2 =.55	51%	48%	0.5
Baseline Pats	θ_1 =2, θ_2 =.65	58%	78%	0.61
New Pats	θ_1 =2, θ_2 =.65	76%	79%	0.77
NASTY	PARAMS	P	R	F
Cue	θ_1 =2, θ_2 =.6	66%	40%	0.58
Baseline Pats	θ_1 =2, θ_2 =.7	86%	55%	0.77
New Pats	θ_1 =2, θ_2 =.7	100%	5%	0.2

Table 1: PARAMS: the best parameters for each feature set P: precision, R: recall, F: weighted f-measure

sarcasm	P	R	F
cue-based	51%	48%	0.5
cue OR patterns	56%	62%	0.57
cue AND patterns	71%	32%	0.57
nasty	P	R	F
cue-based	66%	40%	0.58
cue OR patterns	75%	44%	0.69
cue AND patterns	88%	31%	0.44

Table 2: ; Compares the Cue-Based Classifier to the Combined Classifier; P: precision, R: recall, F: f-measure

plateaued around 20 annotators and is about 90%agreement with 10 annotators, showing that the Mechanical Turk task is well formed and there is high agreement. We begin to form a sarcastic and nasty vocabulary from these cues.

Cue based classifier. We use a development set to measure "goodness" of a cue that could serve as a high precision cue by using the percent sarcastic (nasty) and frequency statistics in the development set. These features rely on how frequent (FREQ) (subject to a θ_1), and how reliable (%SARC and %NASTY) (subject to a θ_2) a cue has to be to be useful. We select candidate cues by exhausting θ_1 = [2, 4, 6, 8, 10] and θ_2 = [.55, .60, .65, .70, .75, .80, .85, .90, .95, 1.00] for $\theta_1 \leq$ FREQ and $\theta_2 \leq$ SARC. At least two cues must be present and above the thresholds in an utterance to be classified by the Cue-Based Classifier. Less than two cues are needed to be classified as the counter-class. We select the best combination of parameters from our training set by selecting the parameters yielding the highest weighted f-measure that favors precision over recall. We then ran the Cue-Based Classifier with the best parameters on a test set. However as previously mentioned, R&W's method expects the Cue-Based Classifier to yield high precision, whereas our results (CUE rows in Table 1) are just barely above baseline.

Pattern Based Classifier. The next step in R&W's method is to create a Pattern-Based Classifier that takes as input the predicted labels from the Cue-Based Classifier. R&W's Pattern-Based Classifier is trained on general, syntactic templates known to exist for subjectivity. These patterns are not limited to exact surface matches as the Cue-Based Classifiers require. We reimplement these patterns, and further developed new patterns specifically fine-tuned towards sarcasm in dialogue. For example, our new pattern `OH RB` (oh adverb) matches utterances like "oh right" and "oh sorry" and the pattern `NP WHphrase` matches "someone who" and "someone what". Patterns are extracted from another development set and we again compute FREQ and %SARC and %NASTY for each pattern subject to $\theta_1 \leq$ FREQ and $\theta_2 \leq$ %SARC or % NASTY. Classifications are made if at least two patterns are present and both are above the specified θ_1 and θ_2, again exhausting all combinations of θ_1 and θ_2. Also following R&W, we do not learn "not sarcastic" or "nice" patterns. The counter-classes are predicted when the utterance contains less than two patterns. We test two Pattern-Based Classifiers: one with the original patterns proposed in R&W (BASELINE PATS) and one with the original patterns in addition to our new, fine-tuned patterns (NEW PATS). Table 1 shows the results of the parameters with the highest weighted f-measure.

The Pattern-Based Classifier performs better on Nasty than Sarcasm. We conclude that R&W's patterns alone generalize well on our Sarcasm and Nasty datasets. By adding the fine-tuned patterns in the NEW PATS Classifier, we see a drastic increase in Sarcasm precision. There seems to be little change in recall for Sarcasm. Furthermore, we see a huge increase in precision for Nasty, but a steep decline in recall with the new patterns. We believe this is because these new patterns are tailored towards sarcastic utterances, not nasty. We did not create our own fine-tuned nasty patterns because we do well with R&W's general patterns.

Combined Classifier. To attempt to create a High-Precision Classifier, we combine the Cue-Based Classifier and the Pattern-Based Classifier. We classify a post as sarcastic if it meets either the criteria of the Cue-Based Classifier (e.g. with $\theta_1 = 2, \theta_2 = .55$ for Sarcasm) or the Pattern-Based Classifier (e.g. with $\theta_1 = 2, \theta_2 = .65$ for Sarcasm). We use the same test set with which we test the Cue-Based Classifier and compare the results (Table 2). We furthermore distinguish between a Combined Classifier that makes a classification if both schemata are true (AND), or if only one is true (OR).

OR does better than only the Cue-Based Classifier for all precision, recall, and f-measure. AND does better for precision by far than the Cue-Based Classifier, but with a lower recall. Despite the very low recall for AND, the f-measure of AND and OR is identical. AND is a more selective classifier, only saying "yes" if both schemata are true. This will naturally yield lower recall, but grant higher confidence in those classified.

We believe our Combined AND Classifier now has a high enough precision to be compared with R&W's first approximation High-Precision, Cue-Based Classifier. After running the Combined Classifier on unannotated data, we select 100 predicted sarcastic and 100 predicted not sarcastic utterances and ask human annotators to label them. We expect a high overlap between annotators and the Combined Classifier, which would indicate that human annotators agree with the labels we are automatically predicting. These results are currently in progress.

Despite the fact that we could not create a first approximation High-Precision, Cue-Based classifier like R&W, we have succeeded in creating a High-Precision Combined Classifier using both cues and fine-tuned patterns (71% precision for sarcasm and 88% precision for nastiness). Future work will involve developing fine-tuned patterns for nastiness and exploring different patterns for sarcasm.

References

G.A. Bryant and J.E. Fox Tree. 2002. Recognizing verbal irony in spontaneous speech. *Metaphor and symbol*, 17(2):99–119.

J.E. Fox Tree and J.C. Schrock. 1999. Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language*, 40(2):280–295.

J. E. Fox Tree. 2010. Discourse markers across speakers and settings. *Language and Linguistics Compass*, 3(1):1–13.

R.W. Gibbs. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1):5–27.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing-Volume 10*, pages 105–112. Association for Computational Linguistics.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics.

Dan Sperber and Deidre Wilson. 1981. Irony and the use-mention distinction. In Peter Cole, editor, *Radical Pragmatics*, pages 295–318. Academic Press, New York.

M. Thelen and E. Riloff. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 214–221. Association for Computational Linguistics.

Marilyn Walker, Pranav Anand, , Robert Abbott, and Jean E. Fox Tree. 2012. A corpus for research on deliberation and debate. In *Language Resources and Evaluation Conference, LREC2012*.