"Sure, I Did The Right Thing": A System for Sarcasm Detection in Speech

Rachel Rakov [1], *Andrew Rosenberg*[2]

[1] Linguistics Department, The Graduate Center CUNY, New York, USA
[2] Computer Science Department, Queens College CUNY, New York, USA

rrakov@gc.cuny.edu, andrew@cs.qc.cuny.edu

Abstract

While a fair amount of work has been done on automatically detecting emotion in human speech, there has been little research on sarcasm detection. Although sarcastic speech acts are inherently subjective, humans have relatively clear intuitions as to what constitutes sarcastic speech. In this paper, we present a system for automatic sarcasm detection. Using a new acted speech corpus that is annotated for sarcastic and sincere speech, we examine a number of features that are indicative of sarcasm. The first set of features looks at a baseline of basic acoustic features that have been found to be helpful in human sarcasm identification. We then present an effective way of modeling and applying prosodic contours to the task of automatic sarcasm detection. This approach applies sequential modeling to categorical representations of pitch and intensity contours obtained via k-means clustering. Using a SimpleLogistic (LogitBoost) classifier, we are able to predict sarcasm with 81.57% accuracy. This result suggests that certain pitch and intensity contours are predictive of sarcastic speech.

Index Terms: sarcasm, prosody modeling, emotion detection, speech recognition

10.21437/Interspeech.2013-239

1. Introduction

Davidov et al. in their 2010 paper describe sarcasm as "A sophisticated form of speech act in which the speakers convey their message in an implicit way". They go on to mention that "an inherent characteristic of sarcasm is that it is frequently difficult to recognize"[1]. It is often the case that in sarcastic speech, the intended pragmatic interpretation is the opposite of the canonical semantic meaning. Much research has been done on how humans recognize and understand sarcastic speech, both in isolation and as distinguished from sincere speech. This research has indicated that sarcasm can be reliably characterized by a number of prosodic cues [2]. However, very little work has been done regarding modeling sarcastic speech for automatic recognition. As speech recognition technology continues to progress forward, it will be important for ASR systems to be able recognize more casual and colloquial speech. As sarcasm may be used to express negative and critical attitudes toward persons or events [2,3], it is very conceivable that ASR systems (particularly in products and spoken dialog systems) that are able to recognize sarcastic speech will be useful in the future. In this paper, we present a system for automatic sarcasm detection. Creating a new acted speech corpus that is annotated for sarcastic and sincere speech, we determine that although sarcasm is an inherently subtle process, it is clearly distinct from sincere speech even in examples presented out-of-context. We then examine a number of features to apply to the task of automatic sarcasm detection. Conventional wisdom suggests that there is a 'sarcastic' tone of voice. A number of people have sought to characterize this quantitatively. Cheang & Pell attempted to identify the possible acoustic cues of sarcastic speech [2]. They identified a number of features that they predicted may be indicative of sarcasm. These features include mean f0, standard deviation of f0, f0 range, mean amplitude, amplitude range, speech rate, harmonics-to-noise ratio (HNR), and one-third octave spectral values (as a measure of nasality). Of these features, they found overall reductions in mean f0, decreases in f0 variation (standard deviation), and changes in HNR to be indicative of sarcastic speech. They then argue that "these findings are most consistent with the idea of an ironic tone of voice [4,5], or more precisely, a sarcastic tone of voice (i.e., the existence of defining prosodic features which are used to communicate sarcasm in speech)." Following from their work, we replicate a number of these features to define our acoustic baseline.

We then present an effective way of modeling and applying prosodic contours to the task of automatic sarcasm detection. This approach focuses on using k-means clustering to determine common patterns of pitch and intensity contours. Following from its success in emotion detection [6], and language identification [7], we use Legendre polynomial expansions to represent prosodic contours. Using the k-means centroids, we create prosodic sequences of pitch and intensity contours which are then used to train n-gram models. Sequential modeling of this sort has had successes in speaker recognition [8,9] and nativeness and genre recognition [10]. The perplexities of these sequences are used as additional features in order to attempt to model contextual prosodic information. We find that these features are predictive of sarcasm.

2. Motivating Background

Tepperman et al. conducted experiments using prosodic, spectral, and contextual cues to automatically identify sarcasm in the phrase 'yeah, right' [11]. They chose to use the Switchboard and Fisher corpora, which primarily consist of telephone conversations between strangers. After testing the accuracy of detection of these cues on both an individual and combined basis, they concluded that prosody on its own it not enough to reliably detect sarcasm, and that a combination of contextual and spectral cues distinguishes sarcasm from sincerity most accurately. However, we find a number of limitations with the work in [11] that inspire and inform the work presented here. Tepperman hypothesizes that there are four types of categorical uses of 'yeah, right' in the speech corpora used – acknowledgement, agreement/disagreement, indirect interpretation, and phrase internal. The authors determined that 'yeah, right', when said sarcastically, only appeared as an indirect interpretation. However, due to the inherent subjective and ambiguous nature of sarcasm, the authors found clear-cut categorization to be difficult to come by: "We found the Switchboard and Fisher examples of sarcastic 'yeah right' often functioned [as evidence of understanding commentary]: not only as humorous

interpretation or commentary, but as a grounding act of sorts, a Request for Acknowledgement or sometimes an Acknowledgement itself." We believe that this lack of clearly defined sarcastic examples may have contributed to their null results.

A second limitation lies in the corpora used for the task. We expect that sarcasm is indicative of informal speech. As such, we expect to see more sarcasm among close friends than between strangers. Under these assumptions, a better choice may have been to use corpora that consisted of recorded speech wherein the conversations were between friends or family. The more formal relationships between speakers in the Switchboard and Fisher corpora may have led to some of the mixed sarcasm instances that are discussed above. This forms part of the motivation to make a new corpus. Additionally, Tepperman et al. hand-annotated this data themselves, not taking into adequate account the inherent subjectivity of the task. Ultimately, the labeling of utterances in isolation was found to be too difficult, and the authors needed to use contextual cues to effectively identify sarcastic productions of 'yeah, right'. When listening to the productions in context, the annotators agreed only 76% of the time. Given the subjectivity of the task, it would have been preferable to have clearer examples of sarcasm that the annotators could agree on in isolation.

As many of our thoughts on the limitations of [11] were focused on corpus problems, it became clear that we would need to create a new annotated corpus. A corpus that contained naturally occurring, clear examples of sarcastic speech would have been ideal. However, in weighing the merits of naturally occurring material and more canonical examples of the phenomena, we opted to use material that is more representative of sarcastic speech. In so doing, we followed motivations of using acted speech in emotion recognition. A fair amount of work in emotion detection has been done using acted speech corpora [12]. One of the reasons that acted speech is appropriate for the task is that it is often an idealized form of the phenomenon[13]. This idealization is realized by acted speech being both more exaggerated and containing of fewer mixed emotions than spontaneous speech. To address the subtlety of sarcasm, we felt the idealized phenomena of acted speech might help us obtain very clear examples of sarcastic speech. While both emotion and sarcasm recognition need to ultimately be applied to naturally occurring as well as acted speech, the understanding of sarcasm is still limited. We hope to increase understanding of the impact of sarcasm on speech production by focusing on this idealized scenario.

3. Materials

3.1. The Daria Sarcasm Corpus

In this section we describe the material we use for our investigation of sarcasm.

3.1.1. Corpus Materials

Our corpus was created from *Daria*, an animated television show that ran on MTV from 1997-2001. We chose to use *Daria* for several reasons. As we discuss in Section 2, prior work involving acted speech corpora has been successful in emotion detection, and as sarcasm detection is a similar task, we hope these successes would carry over. While *Daria* is "acted" in the sense that the speech is produced by voice actors reading a script, it differs from the traditional "acted speech" that has previously been used in emotion research. Where "acted speech" implies that a speaker read an utterance with a prescribed emotion, the pragmatic and paralinguistic content of the *Daria* material is dictated by the character's state and narrative context. We believe this should lead to a more natural expression of sarcasm than traditional "acted speech." Additionally, since the television show uses a stylized animation style, it is difficult to determine sarcasm from facial expression alone. We predicted that this would result in even more exaggerated acted speech than a live-action sitcom could have yielded. Furthermore, as a scripted sitcom, *Daria* leans heavily on sarcasm as a comedic device. This causes the source material to be rich in examples that can be used for investigation.

In obtaining the speech for the corpus, we chose to use dialogue exclusively from the titular character. 150 sentences were extracted from all five seasons of the show. These sentences were extracted from DVDs of the show as avi files, which were then converted to wav files and intensity was normalized using Adobe Audition. We collected what we determined to be 75 sarcastic sentences and 75 sincere sentences – these judgments took context into consideration.

3.1.2. The Perceptions of Sarcastic Speech Survey

In order to generate a gold standard of labels for the data, we ran the Perceptions of Sarcastic Speech Survey in the fall of 2012. The survey, hosted by the site SurveyGizmo, required participants to listen to the 150 sentences and label them as sarcastic or sincere as a forced-choice. Response time was also measured. Participants were able to replay utterances as many times as they wanted, and presentation order was randomized so as to avoid bias. Participants were not given any definition of sarcasm beforehand. This was a deliberate choice; as there are many definitions of sarcasm, rather than influencing the subjects by any one definition we allow subjects to employ their own definition. Thus the ratings we obtain represent the conventional wisdom of what is "sarcastic". Participants were required to be adults with no reported hearing loss. Both native and non-native English speakers were allowed to participate; however, non-native speakers had to have been studying English for at least 3 consecutive years. The survey took approximately 20 minutes to complete. The survey was open from August 2012 through October 2012. 165 participants completed the survey, 149 of them native English speakers, and 16 of them non-native speakers.

3.1.3. Results

Because the participants were asked to make a binary decision on the classification of each sentence (sarcastic or sincere), we expected that the results of the survey would be relatively bimodal. Figure 1 shows the results of the distribution. Although we were expecting a bimodal distribution of the results, we observe a trimodal distribution. We note that this trimodality is observable regardless of the bin size of the histogram. To test this trimodality, we fit several Gaussian Mixture Models (GMMs) to our data and calculated both the AIC and BIC at a variety number of components. We find that we achieve the smallest AIC and BIC when using 3 mixture components. While participants labeled the majority of the sentences as sarcastic or sincere, there are also a substantial number of sentences for which participants were in consistent

disagreement about how the sentences should be labeled. This lead to three groupings of the stimuli: those consistently labeled as sarcastic, those consistently labeled as sincere, and those where there is consistent disagreement. This may be evidence of a thresholding disparity among participants. Since sarcasm is subjective, while there are some sentences that are unambiguously sarcastic, it appears that there are also sentences that are "maybe sarcastic" which people threshold in different ways. Figure 1 also shows that there is more consistent agreement on what is sarcastic than what is sincere. This may also be evidence of a possible uneven distribution of sarcastic and sincere sentences, or of a priming effect – by asking a question about sarcasm, we may have encouraged subjects to be more sensitive to it.

Due to this trimodal distribution of the data, we used a trimodal split in order to refine the corpus for the sarcasm recognition task. We looked at what percentage of participants agreed on labels in order to decide which labels sentences should have. Anything that achieved a "Sarcasm" label with 30% agreement or less was labeled as sincere. Anything that was labeled as "Sarcastic" with 72% agreement higher was labeled as sarcastic. These were determined by the minima of the histogram. This left us with 112 sentences. The data that fell in the middle was excluded from the analyses reported in this paper.

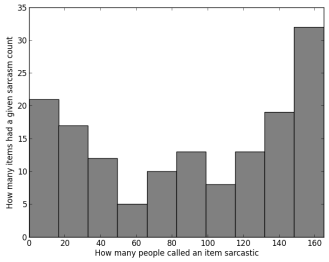


Figure 1: *Histogram showing how many items were labeled as sarcastic and how many people called an item sarcastic.*

4. Methods

4.1. Features

In this section, we describe the acoustic/prosodic features that we investigate as predictive of sarcasm. All acoustic analysis was performed using Snack, a toolkit for Python[14]. Pitch was extracted using the Snack implementation of the ESPS algorithm. Intensity was extracted and converted to decibels.

4.1.1. Sentence level features

The specific acoustic measures derived for each utterance are as follows:

a) mean pitch – measured in log Hertz

b) pitch range – after extracting the top and bottom 5.5% (to avoid outliers), we subtract the minimum pitch from the maximum pitch of the utterance, as a measure of log f0 variation.

c) standard deviation of pitch

d) mean intensity – measured in decibels over the utterance as a whole

e) intensity range – same as pitch range (calculated after removing the top and bottom 5.5% and subtracting the minimum intensity from the maximum intensity of the whole utterance), as a measure of range variation

f) speaking rate – calculated as syllables per log second (the syllable count for each utterance was calculated by counting the canonical number of syllables for each word using cmudict [15]. There are six words in the corpus that do not appear in cmudict. These words were hand transcribed.)

4.1.2. Word level features

Using Praat, all sentences in the corpus were manually word-boundary annotated. We used these word boundaries in order to model prosodic contours within each word. Pitch and intensity contours were modeled using 2-degree i.e. 3 coefficient Legendre polynomial expansions. Doumouchel et al. write, "this approximation of prosodic contour has been used successfully in a number of tasks" [6].

We then clustered the Legendre coefficients of the pitch and intensity contours using scipy's [16] k-means clustering algorithm into 3 distinct groups, respectively. We experimented with k=4 and k=5, but the clusters that resulted from that were less well defined, and yielded worse results in tuning experiments (cf. Section 5).

We then use the resulting centroids of the three clusters to model sequences of prosodic contours. We calculated the Euclidean distance between word level contours and the centroids assigned the label A, B, or C, based on the closest centroid. Using these labels, we were able to construct pitch and intensity contour sequences over sentences at the word-level. By representing prosody as a sequence of word-level symbols, we hope to eliminate unimportant acoustic variation, while maintaining a representation of the suprasegmental prosodic content.

The resulting curves and corresponding Legendre coefficients of the three centroids are presented in Figures 2 and 3.

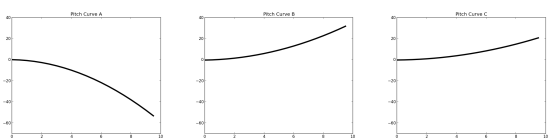
With these sequences in place, we are able to train a sequence model over these prosodic symbols. We explored unigram and bigram modeling. When modeling the unigram sequences, we calculated the percentage of each curve across the sentence as a whole. The bigram model was trained on the entire training corpus. We then calculated the perplexity of each sentence in the train and test corpus against the bigram model. The resulting features are as follows:

g) pitch unigrams A, B, and C – percent of the sentence that is modeled by word level pitch contours A, B, and C,

h) intensity unigrams A, B, and C – percent of the sentence that is modeled by word level intensity contours A, B, and C, ,

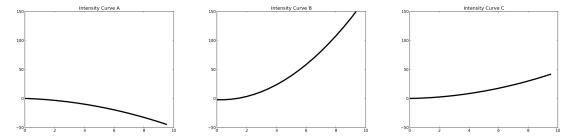
i) pitch bigram perplexity under both the sarcasm and sincere models

j) intensity bigram perplexity under both models



A:[-0.18,-0.26,-0.38] B:[-0.27,0.24,0.22] C:[-0.23,0.13, 0.15]

Figure 2: *Pitch contours*



A:[0.01,-0.88,-0.27] B:[-1.55,-0.81,1.22] C:[0.17,0.47, 0.28]

Figure 3: *Intensity Contours*

We find that when we cluster the pitch contours by an order-2 Legendre polynomial the contours correspond to A) falling pitch, B) a sharper pitch rise and C) a shallower pitch rise. The intensity contours are clustered around the following descriptive patterns: A) shallowly falling intensity, B) sharply rising intensity and C) shallowly rising intensity.

5. Results & Discussion

The corpus was randomly split into a training set and a testing set with 2/3rds of the data used as training. We performed 10-fold cross validation on the train set in order to tune our features and determine which classifier to use. During this tuning process we also experimented with k=4 and k=5 k-means clustering; however, these numbers of clusters did not outperform k=3. Based on the output of this tuning work, we decided to use Weka's SimpleLogistic classifier, a LogitBoost implementation [17] for the classifier for our test data. We use sentence-level acoustic features as a baseline system, and our prosodic modeling features as additional word-level features. Table 1 shows the results on the test set.

In order to evaluate which features were most helpful for the classifier, we used Weka's InfoGainAttributeEval evaluator. These results are reported in Table 2.

5.1.1. Feature set	5.1.2. Percent accuracy
Majority baseline	55.26
Baseline (acoustic features)	65.78
+ pitch unigrams	76.31
+ intensity unigrams	76.31
+ all unigrams	78.94
+unigrams+ pitch bigrams	76.31
+unigram+ intenbigrams	**81.57**
+unigrams + all bigrams	76.31

Table 1: *Results of experiments w/ SimpleLogistic*

Predictive Features	InfoGain
pitch range	0.20
pitch unigram a	0.17
pitch unigram c	0.12
intensity unigram c	0.09

Table 2: *Predictive Features*

Our best results come from a combination of the baseline, unigram counts, and intensity bigram sequence features.

Table 2 lists predictive features, as well as how helpful they are. The most predictive feature from our acoustic baseline was pitch range. We find that sarcastic sentences contain a much reduced pitch range. This is somewhat consistent with what the authors found in [2]. They reported that pitch range was reduced for sarcastic sentences relative to sincere sentences; yet, they only found this to be the case for particular exemplar keyphrase sentences. Our result is consistent with their comment that "changes in the extent of f0 variation produced by speakers is a relatively consistent feature of sarcastic speech, although the direction of these changes is not always uniform and/or this cue may interact more extensively with the nature of the language content [18]."

Regarding prosodic modeling, it is clear that modeling these contours improves sarcasm recognition. When we look at pitch contours, we find that there are fewer instances of falling pitch (A) and shallow pitch rise (C) in sarcastic speech than there are in sincere speech. When we look at intensity contours, we find that sarcastic speech has more instances of shallowly rising intensity (contour C) than sincere speech. This is in keeping with our aforementioned intuitions about sarcastic speech; since sarcasm is a subtle process, abrupt shifts in intensity seem intuitively unlikely. We expect more dynamic intensity in high arousal emotions such as excitement, anger, etc.

The inclusion of our intensity bigram features present some interesting results. While Weka's SimpleLogistic classifier has accesses the intensity bigram features for training, they are ultimately pruned out due to them not be predictive in isolation. However, although the intensity bigram features are not by themselves predictive of sarcasm, when used in conjunction with other features, they encourage these other features to be more effective.

Incorporating the word-level prosodic representation, we achieve an accuracy of 81.57%, a 46.14% relative reduction of error over the sentence-level acoustic baseline.

6. Conclusion

In this paper, we present a system for automatic sarcasm detection. Using a new acted speech corpus that is annotated for sarcastic and sincere speech, we examine a set of both acoustic and prosodic features which may be indicative of sarcasm. We present an effective way of modeling and applying prosodic contours to the task of automatic sarcasm detection. This approach applies sequential modeling of categorical representations of pitch and intensity contours obtained via k-means clustering. Using a SimpleLogistic, LogitBoost classifier, we are able to predict sarcasm with 81.57% accuracy. Our results suggest that certain pitch and intensity contours are predictive of sarcastic speech. This paper presents great potential for future work. One extension would be expanding the application to handle spontaneously occurring sarcasm. More applications include adding additional speakers to the corpus, to see how these same set of features work to identify sarcasm across speakers and languages. Our next task, however, will be to look at the performance of this classification approach on the material that human annotators found to be ambiguous.

7. Acknowledgements

Thanks to Seth Madlon-Kay for his helpful comments. This work was partially supported by DARPA FA8750-13-2-0041 under the Deep Exploration and Filtering of Text (DEFT) Program.

8. References

[1] Davidov, D., Tsur, O., and Rappoport, A. "Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon", Proceedings of the Fourteenth Conference on Computational Natural Language Learning, 107-116, 2010.

[2] Cheang, H.S. and Pell, M.D. "The sound of sarcasm", Speech Commun. 50, 366-381, 2008.

[3] Kreuz, R. J. and Glucksberg, S. "How to be sarcastic: the echoic reminder theory of verbal irony", Journal of Experimental Psychology: General, 118(4), 374-386, 1989.

[4] Clark, H.H., and Gerrig, R.J. "On the pretense theory of irony", Journal of Experimental Psychology: General, 113(1), 121-126.

[5] Mueke, D.C. "The Compass of Irony", Methuen, London, 1969.

[6] Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., and Boufaden, N. "Cepstral and Long-Term Features for Emotion Recognition", Proc. Interspeech Brighton, 344-347, 2009.

[7] Lin, C.Y. and Wang, H.C. "Language Identification Using Pitch Contour Information", ICASSP, 601-604, 2005.thing

[8] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., Stolcke, A. "Modeling prosodic feature sequences for speaker recognition", Speech Commun. 46, 455-472, 2005.

[9] Adami, A.G., Mihaescu R., Reynolds, D., Godfrey, J.J. "Modeling Prosodic Dynamics for Speaker Recogntion", Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[10] Rosenberg, A. "Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness", Proceedings of Interspeech, 2011.

[11] Tepperman, J., Traum, D., and Narayanan, S. "'Yeah Right': Sarcasm Recognition for Spoken Dialogue Systems", Proceedings of InterSpeech ICSLP, 2006.

[12] Vogt, T., Andre, E., and Wagner, J. "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation", Affect and Emotion in Human-Computer Interaction, 75-91, 2008.

[13] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. "A database of german emotional speech", Proceedings of Interspeech, Lisbon, Portugal 2005.

[14] Boersma, P. and Weenink, D. "Praat: Doing Phonetics by Computer(version 5.3.42), 2013. Available: http://www.praat.org

[15] The Carnegie Mellon Pronouncing Dictionary [cmudict.0.7a] https://cmusphinx.svn.sourceforge.net/svnroot/cmusphinx/trunk/cmudict/ Copyright 1993-2008 Carnegie Mellon University

[16] Jones, E., Oliphant, T., Peterson, P. "SciPy: Open source tools for Python", 2001, http://scipy.org

[17] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Whitten, I. H. "The WEKA Data Mining Software: An Update", SIGKDD, 11, 2009.

[18] Bryant, G.A. "Prosodic contrasts in ironic speech", Discourse Processes, 47, 545-566, 2010.