"YEAH RIGHT": SARCASM RECOGNITION FOR SPOKEN DIALOGUE SYSTEMS

Joseph Tepperman[1], *David Traum*[2], *and Shrikanth Narayanan*[1]

[1] Signal Analysis and Interpretation Laboratory, University of Southern California
[2] Institute for Creative Technologies, University of Southern California
tepperma@usc.edu, traum@ict.usc.edu, shri@sipi.usc.edu

10.21437/Interspeech.2006-507

ABSTRACT

The robust understanding of sarcasm in a spoken dialogue system requires a reformulation of the dialogue manager's basic assumptions behind, for example, user behavior and grounding strategies. But automatically detecting a sarcastic tone of voice is not a simple matter. This paper presents some experiments toward sarcasm recognition using prosodic, spectral, and contextual cues. Our results demonstrate that spectral and contextual features can be used to detect sarcasm as well as a human annotator would, and confirm a long-held claim in the field of psychology – that prosody alone is not sufficient to discern whether a speaker is being sarcastic.

Index Terms: dialogue, user modeling, sarcasm, speech acts

1. INTRODUCTION

Sarcasm, also called verbal irony, is the name given to speech bearing a semantic interpretation exactly opposite to its literal meaning. Often used to express humorous analysis or commentary, sarcasm differs from lying in that there is no intent to deceive on the speaker's part [4]. Like a Grounding Act, sarcastic speech requires collaboration – its effect is dependent on the assumed intentions and mutual beliefs held by all dialogue participants [1].

In the realm of spoken dialogue systems, the use of sarcasm presents an interesting fly in the ointment. Because of its complex semantic nature, sarcasm demands that we rethink our fundamental assumptions behind user modeling, grounding strategies, and other methods of automatic spoken dialogue understanding [8]. Sarcastic speech violates at least one of Grice's maxims for cooperative dialogue ("Do not say what you believe to be false") and probably many other pragmatic constraints of any standard spoken dialogue model [2]. Sarcasm also occurs sufficiently frequently in casual conversation that some meeting recognition tasks have purposely chosen to account for it [3]. These reasons all suggest that sarcasm is an important area of inquiry to those working to formulate robust spoken dialogue models.

Here we have restricted our investigation to the expression "yeah right" because of its succinctness as well as its common usage (both sarcastically and otherwise) in conversational American English. Though they are not annotated for sarcasm, the Switchboard and Fisher corpora of spontaneous two-party telephone dialogues contain several hundred occurrences of this expression. Moreover, when used sarcastically "yeah right" is an example of what [5] says to be the most common type of sarcasm – one that has a positive literal meaning but a negative semantic value – and so may serve as an exemplar for other sarcastic forms.

In this study we trained an automatic sarcasm recognizer using spectral, contextual, and prosodic cues. In evaluating the performance and approach to training such a classifier, we intended to learn what types of features are most indicative of sarcasm's presence, formulate how sarcasm functions as a speech act, and recommend how a dialogue manager should respond to sarcasm, once detected.

2. CONTEXTUAL FEATURES

2.1. Sarcasm as a Speech Act

Inspired by work in the area of Grounding Acts [8], and from careful analysis of the data, we chose to categorize each of our examples of "yeah right" as one of the following four types of speech acts (with examples from Switchboard):

Acknowledgment – When a speaker uses a sincere "yeah right" by way of presenting evidence of understanding (in the Grounding Acts sense), we regarded it as an Acknowledgment.
A: Oh, well that's right near Piedmont. (1)
B: Yeah right, right...

Agreement/Disagreement – This act is largely self-explanatory and differs from Acknowledgment in that it's not used to ground understanding of information previously presented. Disagreement, of course, could only occur in the sarcastic case.
A: A thorn in my side: bureaucratics. (2)
B: Yeah right, I agree.

	Laughter	*Q/A*	*Start*	*End*	*Pause*	*Male*
Sarcastic	0.73	0.10	0.57	0.43	0.07	0.23
Sincere	0.22	0.20	0.94	0.48	0.19	0.51

Table 1. Probability of each contextual feature given the class of speech.

Indirect Interpretation – This is for the type of speech act not directed at the dialogue partner but at a hearer not present, as in recounting a past use of a sarcastic "yeah right" by way of telling a story (Switchboard and Fisher actually had several examples of this).

A: "…We have too many pets!" I thought, "Yeah right, come tell me about it!" You know?
B: [laughter] (3)

Phrase-Internal – This is when the "yeah right" does not function as a singular speech unit unto itself, but is included as part of some larger speech act.

A: Park Plaza, Park Suites? (4)
B: Park Suites, yeah right across the street, yeah.

This example may contain the adjacent words "yeah right" but the "yeah" and the "right" are disjunct, the "right" being more properly semantically paired with the word that follows it: "across."

After hand-coding all Switchboard and Fisher occurrences, we noted that there were no sarcastic Acknowledgment or Phrase-Internal realizations, nor did we ever classify a sincere "yeah right" as Indirect Interpretation, so this built-in mutual exclusivity indicated that our chosen act taxonomy did lend itself well to the sarcasm task. But the inherent subjectivity of the coding scheme didn't allow for necessarily clear-cut categorizations. Disambiguating an Agreement from an Acknowledgment could not always be easily or objectively done. Consider the following example:

A: They were never allowed in the house, you know…
B: Uh-huh. Yeah, it would ruin them. [laughter]
A: Yeah, right. [laughter] That's what he said. (5)
B: They say that it'll ruin them.

Speaker A's "Yeah, right" is used as sarcastic interpretation in agreement with B's previous sarcastic comment, but A's use of sarcasm also functions as evidence of understanding B's commentary. We found the Switchboard and Fisher examples of sarcastic "yeah right" often functioned in this way: not only as humorous interpretation or commentary, but as a grounding act of sorts, a Request for Acknowledgment on the part of the hearer, or sometimes an Acknowledgment itself. This Acknowledgment usually came in the form of laughter or explanation of the joke (both of which we see in the above example) – at any rate, some demonstration of evidence of understanding the speaker's sarcastic commentary.

2.2. Objective Cues

To objectively define intersecting marks of sarcasm's (or sincerity's) context and allow for a softer distinction among the speech acts, we used the following feature set:

Laughter – An obvious choice, given sarcasm's often humorous but benevolent intentions. This included laughter either in the "yeah right" turn itself, or in the adjacent turns uttered by either dialogue participant.

Question/Answer – Defining what constitutes an Acknowledgment might not be clear-cut, but what's related yet less subjective was the use of "yeah right" as the answer to a question, which seemed correlated with sincerity, at least in many of the Switchboard and Fisher examples.

Start, End – Did the "yeah right" come at the start or end of the speaker's turn? Or was it perhaps an entire turn itself, serving as both start and end? In the latter case, we found sarcasm to be very unlikely: usually a sarcastic "yeah right" was followed by some elaboration on the point, some explanation of the joke. That's why these seemed like meaningful features.

Pause – In these examples sarcasm was often used as part of a witty repartee, a quick back-and-forth type of dialogue with funny rejoinders again and again. If a "yeah right" was either preceded or followed by a pause in the conversation (even a filled one), that seemed like a clear indication that it could not have been intended sarcastically. Here we usually defined a pause as a lull of longer than 0.5 seconds, though this also depended on the pace of the conversation.

Gender – [5] claims that sarcasm is used more often by men than women. Therefore we included this among the binary contextual features.

In example (5), for instance, the speaker was female, the "yeah right" did not come as the answer to a question, it fell at the beginning (but not end) of the turn, it was preceded and followed by laughter, and there was no pause in the conversation.

3. PROSODIC AND SPECTRAL FEATURES

The following list of 19 prosodic features were chosen to characterize the relative "musical" qualities of each of the words "yeah" and "right" as a function of the whole utterance, by way of parameterizing a perceived sarcastic tone of voice in terms of word pitch, energy, and duration, and other features derived from them. They are similar to those features previously used for detection of non-native speech [6] and syllable stress [7].

1, 2: Average pitch in "yeah" and "right" individually, normalized by the average pitch over the whole utterance
3, 4: Duration of each word, normalized by the utterance's duration
5, 6: Average energy in each word, normalized by the average energy of the utterance
7, 8, 9, 10: The number of rising and falling frames (positive and negative pitch slopes) in each word, normalized by the total frames in that word
11, 12: The number of inter-frame changes from rising to falling pitch slope within each word, also normalized by the number of frames in that word
13, 14: The overall pitch slope from the first to last frame of each word, normalized by the whole utterance's overall pitch range
15: The overall pitch slope of the whole utterance, normalized by the whole utterance's pitch range
16, 17: The pitch range for each word, normalized by the pitch range of the whole utterance
18, 19: The energy range for each word, normalized by the energy range of the whole utterance

Each context-independent recording of "yeah right" (just the "yeah right" without any other words) was automatically segmented on the word level using the "align" fuction of the University of Colorado's Sonic Continuous Speech Recognizer. Then pitch and energy information was extracted using the Snack Sound Toolkit's "pitch" function.

As for spectral features, the first 12 MFCC's (plus Energy) and their delta and acceleration coefficients were extracted from the context-free recordings every 10 msec using a window length of 20 msec, and these were used to train two five-state Hidden Markov Models using embedded re-estimation in HTK. We also tried single-state GMMs, but the HMMs performed better. Because of the sparsity of data, this model training and decoding was done with a leave-one-out crossvalidation procedure. After decoding, we obtained log-likelihood scores representing the probability that our acoustic observations were drawn from each class, sarcastic and sincere. Lastly, these two likelihood scores and their ratio were then used as features (alongside the prosodic and contextual ones) in the final decision-tree-based classification algorithm.

4. DATA ANNOTATION

For purposes of annotation and feature extraction, two versions of the Switchboard and Fisher "yeah right" examples were generated. One was without the surrounding dialogue context, for purely prosody-based classification and annotation. Two human labelers annotated a statistically significant subset of these context-independent recordings for use as a baseline for comparison with an automatic classifier that uses only prosodic features. Inter-annotator agreement was very low: 52.73% (just above the chance agreement of 43.93% - not 50% because the coders did not use both classes in equal proportions), and our Kappa statistic was also very low: 0.1569. But according to [4] and [5] this is to be expected. Both sources argue that prosody alone is not a very reliable indicator of sarcasm. Insofar as a sarcastic tone of voice exists, a listener also relies heavily on contextual and, when available, visual information to identify sarcasm. [5] even goes so far as to claim that the only sarcasm-dependent prosodic feature is energy (or, the speaker's volume). One explanation that may account for this is that perhaps the phrase "yeah right" is used sarcastically so often that, even when it's not meant sarcastically, it might out of habit be spoken in that tone of voice, thus obscuring the true semantic content. At any rate, our two labelers had a hard time telling the difference without any context.

The other version of the data included the two or three dialogue turns before and after the "yeah right," so that each example's "true" class (sarcastic or sincere) might be ascertained. [5] reports that well-trained human labelers, given visual, prosodic, and contextual cues, can agree on what constitutes sarcasm about 80% of the time. Given the surrounding context, our two annotators agreed 76.67% of the time (above the new chance agreement of 66%) and with a Kappa statistic of 0.313, signifying fair correlation. Any disagreements between the two annotators were resolved and these labels were taken as the true classes for the training and test set examples. Similarly, one annotator's own coding of the binary contextual features we took to be a sort of "ground truth" for classifier training and evaluation. There was no need to measure inter-annotator agreement on these because the features themselves were beyond dispute by design, and could be detected automatically fairly easily.

5. DATA STATISTICS

Our dataset comprised a total of 131 uninterrupted occurrences of the phrase "yeah right" found in the Switchboard and Fisher corpora, 30 of which were used sarcastically (about 23%). Though this might seem like a very small percentage, consider that [5] was only able to find 48 sarcastic utterances (not limited to "yeah right") after watching a total of 128 half-hour talk shows. Also, the Switchboard and Fisher recordings are of conversations between strangers, and [5] claims that sarcasm is much more common among close friends, understandably.

The proportions of each contextual feature given the class are enumerated in Table 1. So, for example, of the 101 sincere utterances of "yeah right," only about 22% were spoken in the presence of laughter, compared to 73% of the 30 sarcastic realizations, which seems to confirm the hypothesis behind using laughter as a contextual feature in the first place. This difference in proportions proved to be significant at the 95% confidence level. However, the Switchboard and Fisher data did not support the assertion in [5] that sarcasm is used more often by men than women. With 95% confidence we found we could accept the

	F-measure	*accuracy*
prosodic	0.00	0.69
spectral	0.35	0.77
contextual	0.52	0.84
prosodic + spectral	0.43	0.76
prosodic + contextual	0.46	0.77
contextual + spectral	**0.70**	**0.87**
all 3	0.68	0.86

Table 2. Classification results, sorted by feature set.

alternative hypothesis that the proportion of sarcastic females exceeded that of males in our examples.

6. EXPERIMENTS AND RESULTS

Table 2 reports classification results based on a leave-one-out crossvalidation test procedure over various combinations of features. This decision tree classifier, trained on the C4.5 algorithm implemented in the Weka toolkit [9], was chosen because of its versatility in combining binary and continuous features.

Just as [5] predicted, and as expected by human agreement, using only prosodic features yielded the worst accuracy, though performing better than chance and better than the human annotators could when listening without context. Adding contextual and spectral features to the prosodics did improve accuracy, but leaving the prosody out and just using contextual and/or spectral features exceeded the inter-human agreement of 80% as reported by [5], and our own annotators' agreement of 76.67%. So it seems that prosodic features are unnecessary, when paired with contextual or spectral cues.

Among the contextual features we found that laughter was the most important contribution. And among the prosodic features, it was the number of rising pitch frames in "yeah," the average energy over "right," and the number of rising pitch frames in "right" that most contributed to this classifier's performance; this seems to support [5]'s result that energy contributes the most to the perception of a "sarcastic tone of voice," but it does suggest that pitch is an important aspect of sarcastic prosody as well. This feature ranking is based on the information gain with respect to the given class, using leave-one-out crossvalidation again.

7. CONCLUSIONS

Designers of a dialogue manager or dialogue transcriber wishing to detect sarcasm would probably be best served to just ignore prosodic features entirely and concentrate on extracting the kinds of contextual features we used in this study. Most of them would be simple enough to detect automatically, either from acoustics (for laughter, gender, pause) or semantics and syntax (question/answer, start/end).

As for handling the sarcasm once it's detected, a dialogue agent ought to do what real humans do and acknowledge it. Either generate some synthetic laughter or, for more advanced agents, somehow point out that it "gets" the joke.

Future work in this area could focus on generalizing these findings for other utterances besides "yeah right"; using a finer-grain taxonomy to accommodate subtle differences in the types of sarcasm (good-natured vs. biting, for example); the incorporation of other, more meaningful prosodic features, as well as visual cues; and, of course, acquiring more data.

8. ACKNOWLEDGMENTS

Thanks to Abe Kazemzadeh and Francesca Giancoli for labeling some of the data.

9. REFERENCES

[1] D. G. Field and A. Ramsay. "Sarcasm, deception and stating the obvious: Planning dialogue without speech acts." *Artificial Intelligence Review* 22(2): 149-177. 2004.

[2] H. P. Grice. "Logic and Conversation." In P. Cole and J. Morgan (eds.), Syntax and Semantics, Vol. 3: Speech Acts, pp. 41-58, New York: Academic Press, 1975.

[3] D. Hillard, M. Ostendorf, and E. Shriberg. "Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data." Proc. HLT-NAACL Conference, Edmonton, Canada, May 2003.

[4] C. Nakassis and C. Snedeker. "Beyond sarcasm: Intonation and Context as Relational Cues in Children's Recognition of Irony." In A. Greenhill, M. Hughs, H. Littlefield, & H. Walsh (eds.), Proceedings of the Twenty-sixth Boston University Conference on Language Development. Somerville, MA: Cascadilla Press. 2002.

[5] P. Rockwell. "Sarcasm on Television Talk Shows: Determining Speaker Intent Through Verbal and Nonverbal Cues." In A. Clark (ed.), Psychology of Moods. New York: Nova. 2005.

[6] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, K. Sonmez, "Evaluation of Speaker's Degree of Nativeness Using Text-Independent Prosodic Features," in *Proc. of the Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[7] J. Tepperman and S. Narayanan, "Automatic Syllable Stress Detection for Pronunciation Evaluation of Language Learners," *Proc. ICASSP'05*, Philadelphia, 2005.

[8] D. Traum and E. Hinkelman, "Conversation Acts in Task-Oriented Spoken Dialogue," In *Computational Intelligence*, 8(3): 575-599, 1992.

[9] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005.