Let's talk (ironically) about the weather: Modeling verbal irony

Justine T. Kao (justinek@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University

Abstract

Verbal irony plays an important role in how we communicate and express our opinions about the world. While there exist many theories and empirical findings about how people use and understand verbal irony, there is to our knowledge no formal model of how people incorporate shared background knowledge and linguistic information to communicate ironically. Here we argue that a computational approach previously shown to model hyperbole (Kao, Wu, Bergen, & Goodman, 2014) can also explain irony once we extend it to a richer space of affective subtext. We then describe two behavioral experiments that examine people's interpretations of utterances in contexts that afford irony. We show that by minimally extending the hyperbole model to account for two dimensions of affect—valence and arousal—our model produces interpretations that closely match humans'. We discuss the implications of our model on informal theories of irony and its relationship to other types of nonliteral language understanding.

Keywords: irony; computational modeling; pragmatics; nonliteral language understanding

Introduction

For better or for worse, verbal irony—defined as utterances whose apparent meanings are opposite in polarity to the speaker's intended meaning (Roberts & Kreuz, 1994; Colston & O'Brien, 2000)—is a major figurative trope of our time. From popular sitcoms to political satire to *#sarcasm* on Twitter and casual conversations among friends, verbal irony plays an important role in how we communicate and express opinions about the world. The prevalence of verbal irony poses a puzzle for theories of language understanding: Why would speakers ever use an utterance to communicate its opposite meaning, and how can listeners appropriately interpret such an utterance? Previous work has shown that verbal irony serves several important communicative goals, such as to heighten or soften criticism (Colston, 1997b), elicit emotional reactions (Leggitt & Gibbs, 2000), highlight group membership (Gibbs, 2000), and express affective attitudes (Colston & Keller, 1998; Colston, 1997a). These findings suggest that while ironic statements are false under their literal meanings, they are often highly informative with respect to social and affective meanings. In this paper we argue that a computational approach previously shown to explain hyperbole (Kao, Wu, et al., 2014) can also model irony—once extended to a richer space of affective subtext.

Linguists and psychologists have proposed several informal theories of how people understand verbal irony. According to a classic Grician analysis, listeners first need to recognize that an ironic utterance blatantly violates the maxim of quality (to be truthful); they then arrive at a conversational implicature that the intended meaning is contrary to the utterance's literal meaning (Grice, 1967; Wilson, 2006). While Grice's account is appealing in its treatment of verbal irony as arising naturally from conversational maxims, it does not provide a detailed explanation for how the appropriate implicature is derived from these maxims, or why it is ever rational to deliver an utterance that is opposite from the truth (Wilson, 2006). Other theorists have responded by invoking echoic mentions of past experiences (Sperber & Wilson, 1981; Jorgensen, Miller, & Sperber, 1984) or the notion that irony is a form of pretense (Clark & Gerrig, 1984). We instead explore the idea that irony will emerge from general principals of communication when properly formalized in a model that accounts for uncertain and potentially affective topics of conversation.

Rational Speech Act (RSA) models are a family of computational models that formalize language understanding as recursive reasoning between speaker and listener, and have been shown to account for many phenomena in pragmatics (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). Kao, Wu, et al. (2014) introduce a critical extension to the basic RSA model by considering the idea that listeners may be uncertain which question under discussion (QUD – or topic of conversation) the speaker aims to address when formulating an utterance. To understand an utterance, the listener jointly infers the QUD as well as the speaker's intended meaning. For example, a speaker may want to communicate negative affect about a situation (e.g. unhappiness about the cool temperature outside) instead of the precise situation (the temperature outside), in which case choosing an exaggerated utterance ("It's freezing outside!") effectively communicates negative affect and addresses the QUD. A listener who reasons about the speaker and QUD is then able to use his background knowledge to correctly infer that the speaker is upset about the temperature, but that it is unlikely to be literally freezing outside (especially if she is in California). Kao, Wu, et al. (2014) showed that this model—which we will refer to as qRSA—produces nonliteral interpretations of hyperbolic utterances that closely match humans'; however, they considered only a simplified affect space, namely the presence or absence of negative feeling. This overlooks the range of attitudes and emotions that speakers could express with nonliteral utterances. In particular, since verbal irony involves expressing negative meanings with positive utterances and vice versa, a richer space of affect that includes both positive and negative emotions may be key. Here we examine the consequence of considering the range of emotions in an empirically derived affect space within the qRSA model; we show that this minimal change is able to capture many of the rich inferences resulting from verbal irony.

In what follows, we will examine interpretations of potentially ironic utterances in an innocuous domain—the weather. We chose the weather as the victim of irony for several reasons. First, people are quite familiar with talking (and com-

plaining) about the weather. Second, we can visually represent the weather to participants with minimal linguistic description in order to obtain measures of nonlinguistic contextual knowledge. Finally, we can vary the weather states to observe how the same utterance is interpreted differently given different contextual knowledge. We first explore how an enriched space of affect impacts the qRSA model and find that it produces ironic interpretations. We then present two behavioral experiments that examine people's interpretations of utterances given different weather contexts. We show that by accounting for two types of affective dimensions, valence and arousal, our model produces interpretations that closely match humans'. Finally, we discuss implications of our model for informal theories of irony and its relationship to other types of nonliteral language understanding.

Computational Model

In this section, we describe the qRSA model[1] and compare different spaces of affect to test the conditions for producing ironic interpretations. Following the qRSA model described in Kao, Wu, et al. (2014), a speaker chooses an utterance that most effectively communicates information regarding the question under discussion (QUD) to a literal listener. We consider a meaning space consisting of the variables s, A, where s is the state of the world, and A represents the speaker's (potentially multidimensional) affect towards the state. We formalize a QUD as a projection from the full meaning space to the subset of interest to the speaker, which could be s or any of the dimensions of A. We specify the speaker's utility as information gained by the listener about the topic of interest—the negative surprisal of the true state under the listener's distribution given an utterance, u, along the QUD dimension, q. This leads to the following utility function:

$$U(u|s,A,q) = \log \sum_{s',A'} \delta_{q(s,A)=q(s',A')} L_{literal}(s',A'|u) \qquad (1)$$

where $L_{literal}$ describes the literal listener, who updates her prior beliefs about s, A by assuming the utterance to be true of s. The speaker S chooses an utterance according to a softmax decision rule (Sutton & Barto, 1998): $S(u|s,A,q) \propto e^{\lambda U(u|s,A,q)}$, where λ is an optimality parameter. A pragmatic listener $L_{pragmatic}$ then takes into account prior knowledge and his internal model of the speaker to determine the state of the world as well as the speaker's affect. Because $L_{pragmatic}$ is uncertain about the QUD, he marginalizes over the possible QUDs under consideration:

$$L_{pragmatic}(s,A|u) \propto P(s)P(A|s)\sum_{q} P(q)S(u|s,A,q)$$

The resulting distribution over world states and speaker affects is an *interpretation* of the utterance.

We performed the following simulations to examine the model's behavior using affect spaces, A, that differ in complexity and structure. We assume that s has five possible ordered values: `terrible`, `bad`, `neutral`, `good`, and `amazing`. We consider two different weather contexts: apparently bad weather and apparently amazing weather, which are each specified by a prior distribution over these states (see gray dotted lines in Figure 1). We then examine how the model interprets the sentence "The weather is terrible" in the two weather contexts, given different affect spaces.

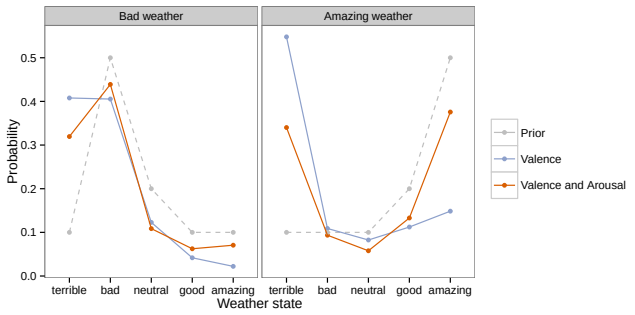

Figure 1: Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and affect dimensions. Gray dotted lines indicate prior beliefs about weather states given a weather context; blue lines indicate interpretations when reasoning only about the speaker's valence; orange lines indicate interpretations when reasoning about both valence and arousal.

We first consider a one-dimensional affect space, where the dimension is emotional valence, and the values are whether the speaker feels negative or positive valence towards the state. The blue lines in Figure 1 show the model's interpretation of "The weather is terrible" using this one-dimensional affect space. The model is capable of non-literal interpretation: it produces a hyperbolic interpretation (that the weather is merely `bad`) given "The weather is terrible" in the bad weather situation. However, it produces a literal interpretation (that the weather is `terrible`) in the amazing weather situation. In other words, a model that only considers valence is unlikely to infer a positive world state from a negative utterance (and vice versa), thus failing to evidence verbal irony.

Affective science identifies two dimensions, termed valence and arousal, as underlying the slew of emotions that people experience (Russell, 1980). For example, *anger* is a negative valence and high arousal emotion, while *contentment* is a positive valence and low arousal emotion. Could speakers leverage the arousal dimension to convey high arousal and positive affect (e.g. excitement) using utterances whose literal meanings are associated with high arousal but negative affect (e.g. "The weather is terrible!")? The orange lines in Figure 1 show simulations of the qRSA model with a two-dimensional affect space: whether the speaker feels negative/positive valence and low/high arousal towards the weather state. Given strong prior belief that the weather state is `bad`, the model interprets "The weather is terrible" to mean that the weather is likely to be `bad`, again producing a hyper-

[1] See Kao, Wu, et al. (2014) for modeling details and a more leisurely exposition.

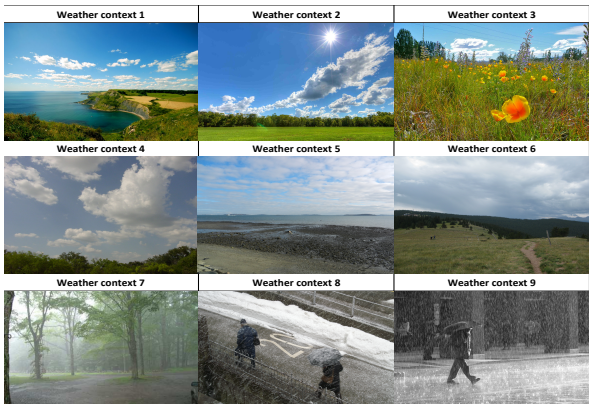

Figure 2: Weather images shown to participants in Experiments 1 and 2.

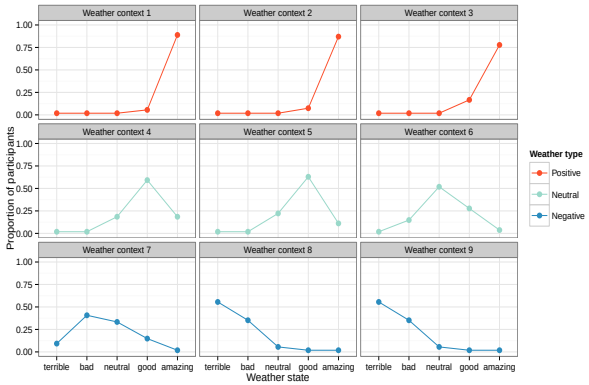

Figure 3: Proportion of participants who rated each weather context as each weather state.

bolic interpretation. However, given strong prior belief that the weather is `amazing`, the model now places much greater probability on the ironical interpretation of "The weather is terrible," meaning that the weather is likely `amazing`. This is because, with the enriched two-dimensional affect space, the pragmatic listener realizes that the speaker may be using "terrible" to communicate high emotional arousal. Note that this result is not simply due to the model falling back on the prior: given the same priors, the model interprets the neutral utterance "The weather is ok" as the weather state being `neutral` and not `amazing`. These simulations suggest that a psychologically realistic, two-dimensional affect space enables the model to appropriately interpret ironic utterances in addition to hyperbolic ones.

Behavioral Experiments

To quantitatively test whether the qRSA model with expanded affect space can capture a range of ironic interpretations, we need appropriate prior distributions as well as data for human interpretations. We conducted Experiment 1 to measure prior beliefs over weather states ($P(s)$) for various weather contexts as well as the likelihood of various emotions towards different weather states. The latter allows us to empirically derive the affective space and priors, $P(A|s)$, for this domain. In Experiment 2, we collected people's ratings of how a speaker perceives and feels about the weather given what she says (e.g. "The weather is terrible!" when the context clearly depicts sunny weather).

Experiment 1: Prior elicitation

Materials and methods We selected nine images from Google Images that depict the weather. To cover a range of weather states, three of the images were of sunny weather, three of cloudy weather, and three of rainy or snowy weather. We refer to these images as weather contexts. Figure 2 shows these nine images. 49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images in random order. In each trial, participants were told that a person (e.g. Ann) looks out the window and sees the view depicted by the image. They then indicated how Ann would rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used slider bars to rate how likely Ann is to feel each of the following seven emotions about the weather: *excited*, *happy*, *content*, *neutral*, *sad*, *disgusted*, and *angry*, which are common emotion categories (Ekman, 1992)[2]. The order of the emotions was randomized for each participant but remained consistent across trials for the same participant. The end points of the slider bars were labeled as "Impossible" and "Absolutely certain." A link to the experiment is here: `http://stanford.edu/~justinek/irony_exp/priors/priors.html`

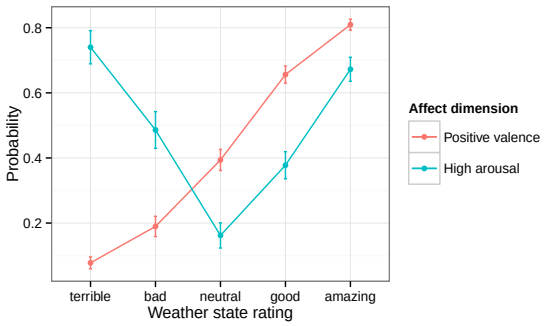

Figure 4: The average probabilities of positive valence and high arousal associated with each weather state. Error bars are 95% confidence intervals. The `terrible` weather state is associated with low probability of positive valence and high probability of high arousal; `amazing` is associated with high probability of positive valence and high probability of high arousal.

[2]From the most frequently cited set of six basic emotions, we removed *fear* and *surprise* and added *content* and *excited* to have a balanced set of positive and negative emotions. We also added *neutral* to span a wider range of emotional arousal.

Results For each of the nine weather contexts, we obtained the number of participants who gave each of the weather state ratings and performed add-one Laplace smoothing on the counts. This allowed us to compute a smoothed prior distribution over weather states given each context, namely $P(s)$. Figure 3 shows that the sunny and positive weather contexts were more likely to be rated as `amazing`, while the negative weather contexts were more likely to be rated as `bad` or `terrible`.

To examine participants' ratings of the affect associated with each context, we first performed Principal Component Analysis (PCA) on the seven emotion category ratings. This allowed us to compress the ratings onto a lower-dimensional space and reveal the main affective dimensions that are important in this domain, as often done in affective science (Russell, 1980). We found that the first two principal components corresponded to the dimensions of emotional valence and emotional arousal, accounting for 69.14% and 13.86% of the variance in the data, respectively. The PCA represents emotion ratings for each trial as real values between negative and positive infinity on each of the dimensions. To map these values onto probability space, we first standardized the scores on each dimension to have zero mean and unit variance. We then used the cumulative distribution function to convert the standardized scores into values between 0 and 1. This gives us the probabilities of Ann feeling positive (vs. negative) valence and high (vs. low) arousal for each trial, which is a two-dimensional probabilistic representation of her affect. By calculating the average probabilities of positive valence and high arousal given each weather state rating, we obtain the probability of positive valence and high arousal associated with each weather state, namely $P(A|s)$ (Figure 4).

Experiment 2: Irony understanding

Results from Experiment 1 give us the components to generate interpretations of utterances from our model. Here we describe an experiment that elicits people's interpretations of utterances, which we then use to evaluate model predictions.

Materials and methods 59 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each participant saw all nine images from Figure 2 in random order. In each trial, participants were told that a person (e.g. Ann) and her friend are in a room looking out the window together and see the view depicted by the image. Ann says, "The weather is _____!" where the adjective is randomly selected at each trial from the following set: "terrible," "bad," "ok," "good," and "amazing." Participants first rated how likely it is that Ann's statement is ironic using a slider with end points labeled "Definitely NOT ironic" and "Definitely ironic." They then indicated how Ann would actually rate the weather using a labeled 5-point Likert scale, ranging from `terrible`, `bad`, `neutral`, `good`, to `amazing`. Finally, participants used sliders to rate how likely it is that Ann feels each of seven emotions about the weather. A link to the experiment is here: `http://stanford.edu/~justinek/irony_exp/interpretation/interpretation_askIrony.html`

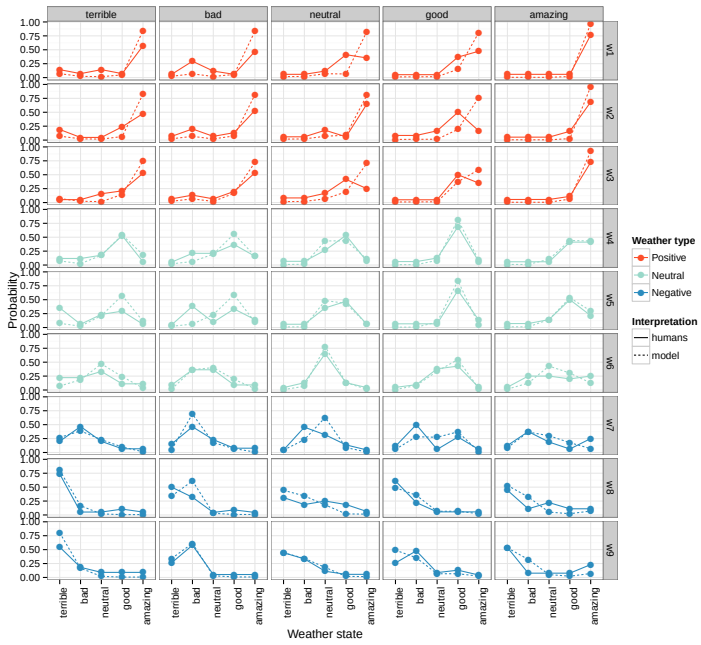

Figure 5: Model's and participants' inferences about the weather state (x-axis) given a weather context (row) and an utterance (column). Each panel represents interpretations of an utterance in a weather context. The solid lines are participants' ratings; the dotted lines are model's posterior distributions over weather states.

Results We first examined participants' irony ratings for each of the weather context and utterance pairs. We found a basic irony effect, where utterances whose polarities are inconsistent with the polarity of the weather context are rated as significantly more ironic than utterances whose polarities are consistent with the weather context ($t(34.16) = -11.12, p < 0.0001$). For example, "The weather is terrible" (a negative utterance) is rated as more ironic in weather context 1 (positive context) ($M = 0.90$, $SD = 0.21$) than in weather context 7 (negative context) ($M = 0.15$, $SD = 0.27$). A linear regression model with the polarity of the utterance, the polarity of the weather context, and their interaction as predictors of irony ratings produced an adjusted R^2 of 0.91, capturing most of the variance in the data. This suggests that participants' lay judgments of irony align with its basic definition: utterances whose apparent meanings are opposite in polarity to the speaker's intended meaning.

Given the fact that participants can identify verbal irony based on its inconsistency with context, how do they then use context to determine the speaker's intended meaning? We examined participants' interpretations of utterances given contexts. For each of the 45 weather context (9) $\times$ utterance (5) pairs, we obtained the number of participants who gave each of the five weather state ratings (`terrible`,

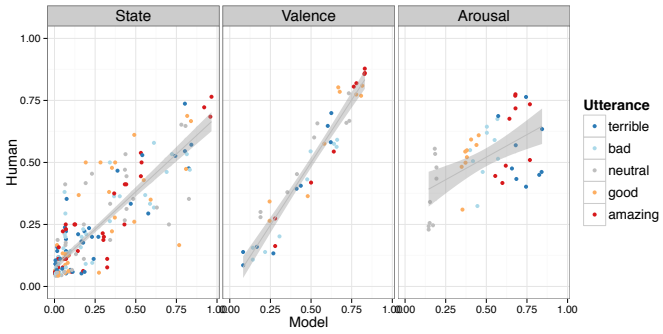

Figure 6: Scatter plot showing correlations between model predictions and human interpretations for weather state, speaker valence, and speaker affect. Colors indicate different utterances.

bad, neutral, good, amazing). We performed add-one Laplace smoothing on the counts to obtain a smoothed distribution over weather states given each context and utterance. The solid lines in Figure 5 show these distributions of ratings. We see that participants produce ironic interpretations of utterances, such that the weather is most likely to be amazing given that the speaker said "The weather is terrible" in weather context 1. Participants also produce hyperbolic interpretations, such that the weather is most likely to be bad given that the speaker said "The weather is terrible" in weather context 7. This suggests that people are highly sensitive to context when interpreting utterances, and use it both to determine when an utterance is not meant literally and to appropriately recover the intended meaning. Finally, we examine participants' inferences about the speaker's affect given utterances in context. We used the loadings from the PCA on emotion ratings from Experiment 1 to project the emotion ratings from Experiment 2 onto the same dimensions. We then standardized and used the cumulative distribution function to convert the scores into values between 0 and 1, as before, which gives us probability ratings of the speaker feeling positive valence and high arousal given an utterance and weather context.

Model Evaluation

We now evaluate the model's performance against these behavioral results. From Experiment 1, we obtained the prior probability of a weather state given a context ($P(s)$) as well as the probability of affect given a weather state ($P(A|s)$). In addition, we fit three free parameters to maximize correlation with data from Experiment 2: the speaker optimality parameter ($\lambda = 1$) and the prior probability of each of the three QUDs ($P(q_{state}) = 0.3$, $P(q_{valence}) = 0.3$, $P(q_{arousal}) = 0.4$)[3]. For each of the 45 utterance and weather context pairs, the model produced an interpretation consisting of the joint posterior distribution $P(s,A|u)$, where A can be further broken down into valence and arousal dimensions. We will examine the model's performance on each of these state and affect dimensions by marginalizing over the other dimensions.

[3]Since $P(q_{state}) + P(q_{valence}) + P(q_{arousal}) = 1$, $P(q_{arousal})$ is determined by the other two QUD parameters and not a free parameter.

Model	State	Valence	Arousal	Average
Literal	0.38	0.45	0.49	0.44
Prior	0.79	0.84	0.49	0.71
Valence	0.84	0.79	0.61	0.75
Valence + arousal	0.86	0.96	0.66	0.83
Best possible	0.90	0.95	0.76	0.87

Table 1: Correlation coefficients between model predictions and human interpretations of weather state, valence, and arousal given an utterance and weather context from Experiment 2. *Best possible* gives an estimate of the maximum possible correlation given noise in the data (see footnote 4).

Figure 6 shows scatter plots correlating model predictions with human interpretation data for each of the dimensions: weather state, valence, and arousal. The model predictions of weather state given utterance match humans' interpretations, with a correlation of 0.86. Since the split-half correlation for the human data is $\rho = 0.898$ (95%CI $= [0.892, 0.903]$)[4] we find that our model captures much of the explainable variance in human judgements. The model predicts humans' interpretations of valence extremely well, with a correlation of 0.96, capturing essentially all of the explainable variance in the data ($\rho = 0.948 \pm 0.001$). Importantly, the model infers the appropriate valence even when it is inconsistent with the valence of the utterance's literal meaning. The model's predictions for emotional arousal match humans' with a correlation of 0.66, capturing a substantial amount of the explainable variance ($\rho = 0.763 \pm 0.005$). Furthermore, the absolute difference between the model's inferred valence and the valence of the utterance's literal meaning correlates significantly with people's irony ratings ($r = 0.86$, $\rho = 0.94 \pm 0.005$), suggesting that the model is able to use inconsistencies between literal and interpreted meanings to identify ironic uses.

We considered a series of simpler models to show that the full model using a two-dimensional affect space best predicts human interpretations. We first examined a model that interprets utterances literally, such that "The weather is terrible" is always interpreted as the weather state being terrible, along with the valence and arousal associated with terrible weather. We then examined a model that simply ignores the speaker's utterance and takes into account only the state and affect priors associated with each weather context. Finally, we examined the performance of the qRSA model with a unidimensional affect space (valence only). Table 1 shows the models' correlations with human judgements for state, valence, and affect. A complete model that takes into account

[4]Split-half correlations ρ were calculated by repeatedly bootstrapping samples from the data (sample each participant with replacement), computing correlation between two halves of the bootstrapped samples, and using the Spearman-Brown prediction formula to predict reliability with full sample size. Confidence intervals are 95% CI over 1000 iterations of bootstrap sampling.

prior knowledge, the literal meaning of the utterance, and a two-dimensional affect space outperforms the other models. This dominance is especially apparent with respect to inferences about valence, which is the most important aspect of understanding an ironic utterance, since the listener must infer the intended positive/negative valence from an ostensibly negative/positive utterance. These comparisons suggest that the model successfully leverages richer knowledge of affect to produce the appropriate nonliteral interpretations.

Discussion

In this paper, we explored the consequences of expanding the space of affect considered by previous Rational Speech Act models to account for verbal irony. We showed that by making a minimal extension to Kao, Wu, et al. (2014)'s hyperbole model, we can capture people's fine-grained interpretations of ironic utterances in addition to hyperbole. This suggests that hyperbole and irony stem from the same underlying principles of communication. We also provided quantitative evidence that irony interpretation relies heavily on context and prior knowledge—this in turn may explain why using irony highlights common ground and group membership (Gibbs, 2000).

There remain important qualities of verbal irony to account for. For example, the echoic mention theory of irony claims that speakers often use verbal irony to remind the listener of previous utterances that turned out to be false or irrelevant, or of positive norms that were violated (Sperber & Wilson, 1981; Jorgensen et al., 1984). On the other hand, pretense theory argues that when a speaker produces an ironic utterance, she is not genuinely making the utterance, but only pretending to be someone who would make such an utterance (Clark & Gerrig, 1984). While our model is able to capture many of the main characteristics of verbal irony, it does not capture the intuitions behind echoic mention or pretense theories. We hope to further improve our model's performance and enrich its understanding of the social aspects of irony by addressing these intuitions in future research.

In addition to shedding light on the communicative principles underlying irony understanding, our work also has interesting connections to natural language processing. Many researchers aim to automatically detect sarcasm in large bodies of texts in order to recover the correct sentiment from an ostensibly positive or negative utterance (Davidov, Tsur, & Rappoport, 2010; Filatova, 2012). A critical insight that emerged from these efforts is that irony detection requires information far beyond surface linguistic cues, often calling upon a deep understanding of context and common knowledge between speaker and listener that computers currently lack (González-Ibáñez, Muresan, & Wacholder, 2011; Wallace, Do Kook Choe, & Charniak, 2014). By integrating background knowledge and linguistic meaning in a principled manner, our model produces ironic interpretations in a way that is highly sensitive to context and common ground.

Overall, our experimental paradigm and modeling framework provide a detailed and precise account of irony understanding. Given the prevalence of irony in everyday language and the social functions it serves, we believe it would be amazing to understand how people interpret utterances that convey the opposite of what they ostensibly mean (#notsarcastic).

References

Clark, H. H., & Gerrig, R. J. (1984). On the pretense theory of irony.

Colston, H. L. (1997a). "i've never seen anything like it": Overstatement, understatement, and irony. *Metaphor and Symbol*, *12*(1), 43–58.

Colston, H. L. (1997b). Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, *23*(1), 25–45.

Colston, H. L., & Keller, S. B. (1998). You'll never believe this: Irony and hyperbole in expressing surprise. *Journal of psycholinguistic research*, *27*(4), 499–513.

Colston, H. L., & O'Brien, J. (2000). Contrast of kind versus contrast of magnitude: The pragmatic accomplishments of irony and hyperbole. *Discourse Processes*, *30*(2), 179–199.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107–116).

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169–200.

Filatova, E. (2012). Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Lrec* (pp. 392–398).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Gibbs, R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, *15*(1-2), 5–27.

González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: short papers-volume 2* (pp. 581–586).

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Grice, H. P. (1967). Logic and conversation. *The Semantics-Pragmatics Boundary in Philosophy*, 47.

Jorgensen, J., Miller, G. A., & Sperber, D. (1984). Test of the mention theory of irony. *Journal of Experimental Psychology: General*, *113*(1), 112.

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th annual meeting of the cognitive science society.*

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Leggitt, J. S., & Gibbs, R. W. (2000). Emotional reactions to verbal irony. *Discourse processes*, *29*(1), 1–24.

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science*, *5*(3), 159–163.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sperber, D., & Wilson, D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, *49*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). Cambridge Univ Press.

Wallace, B. C., Do Kook Choe, L. K., & Charniak, E. (2014). Humans require context to infer ironic intent (so computers probably do, too). *ACL*.

Wilson, D. (2006). The pragmatics of verbal irony: Echo or pretence? *Lingua*, *116*(10), 1722–1743.