Let's talk (ironically) about the weather: A computational model of verbal irony

Justine T. Kao (justinek@stanford.edu)
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Noah D. Goodman (ngoodman@stanford.edu)
Department of Psychology, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Verbal irony plays an important role in how we communicate and express our opinions about the world. While there exist many interesting 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 present two behavioral experiments that examine people's interpretations of utterances given different contexts. We then describe a computational model that reasons about background knowledge, affect, and the speaker's communicate goals to interpret ironic utterances and their rich affective subtexts. We show that by accounting for two types of affect goals—valence and arousal—our model produces interpretations that closely match humans'. Finally, we discuss the implications of our model on 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 present a computational model and behavioral experiments to show that people may use inferences about these alternative dimensions of meaning and the speaker's communicative goals to understand ironic utterances.

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 or satisfactory 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). On the other hand, previous work suggests that it is possible and indeed desirable to consider nonliteral language understanding as a product of general principals of communication, such as reasoning about informativeness with respect to the speaker's communicative goals (Kao, Wu, Bergen, & Goodman, 2014; Kao, Bergen, & Goodman, 2014). In particular, a model that reasons about the speaker's affect is able to interpret hyperbolic utterances and infer the appropriate affective subtext (Kao, Wu, et al., 2014). Given the fact that many researchers believe hyperbole and irony to be closely related phenomena (cite), we suggest that similar principles may account for verbal irony understanding as well. Our goal in this paper is to identify these communicative principles and provide a precise formal account of how they interact to produce ironic interpretations.

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) introduces a critical extension to basic RSA models by considering the idea that speakers may aim to address different questions under discussion (QUDs) when formulating an utterance. An important task for the listener is to then jointly infer 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 temperature outside) instead of the precise situation (e.g. the temperature outside), in which case choosing an exaggerated utterance (e.g. "It's freezing outside!") effectively addresses the QUD. A listener who reasons about the speaker and QUD is then able to use his background knowledge about temperatures 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). Extending this model to consider QUDs opens up the possibility for a speaker to produce an utterance that is literally false but satisfies her goal to

convey affect. While this model—which we will refer to as qRSA—produces nonliteral interpretations of hyperbolic utterances that closely match humans', Kao, Wu, et al. (2014) considered only a unidimensional space of affects, 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 is necessary. Here we examine the consequence of considering a range of emotions in an empirically derived affect space within the qRSA model, and 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 complaining) 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 describe an extension to the qRSA model and show that an enriched space of affect enables the model to produce 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 affect goals, valence and arousal, our model produces interpretations that closely match humans'. Finally, we discuss the implications of our model on irony and its relationship to other types of nonliteral language understanding.

Computational Model

In this section, we describe the qRSA model 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 define the speaker's utility as the negative surprisal of the true state under the listener's distribution given an utterance along the QUD dimension, leading to the following utility function[1]:

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

[1] See Kao, Wu, et al. (2014) for details.

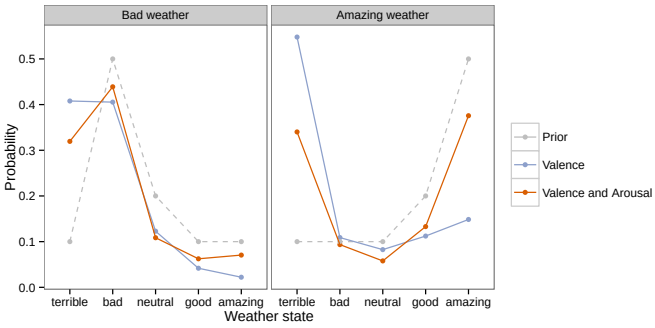


Figure 1: Model interpretations of "The weather is terrible" given different prior beliefs about the weather state and different affect dimensions. The left panel represents strong prior belief that the weather is `bad`; the right panel represents strong prior belief that the weather is `amazing`. Gray dotted lines indicate prior beliefs about weather states; blue lines indicate interpretations when reasoning only about the speaker's valence; orange lines indicate interpretations when reasoning about both valence and arousal. The model requires both valence and arousal to produce an ironic interpretation of the utterance given `amazing` weather.

where q is the QUD and $L_{literal}$ is the literal listener. The speaker S chooses an utterance according to a softmax decision rule (?, ?):

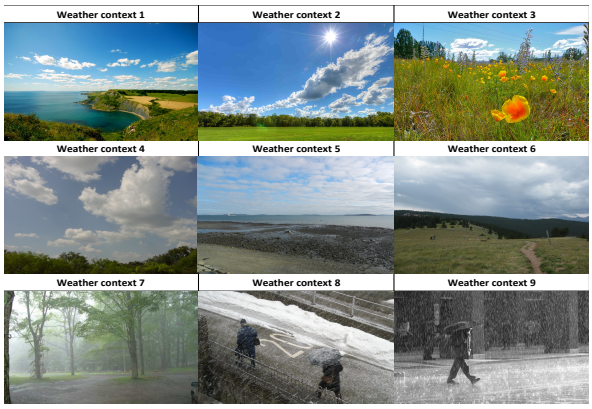
$$S(u|s,a,q) \propto e^{\lambda U(u|s,a,q)}, \qquad (2)$$

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:

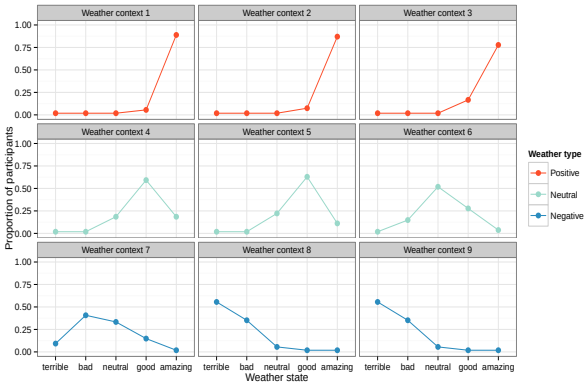
$$L_{pragmatic}(s,a|u) \propto P(s)P(a|s)\sum_{q} P(q)S(u|s,a,q)$$

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

We examine the model's behavior using affect spaces with different dimensions. We first consider a one-dimensional affect space, where the dimension is emotional valence—whether the speaker feels negative or positive about the world state. Suppose there is strong prior belief that the weather state is `bad`; also suppose that the QUD is the speaker S's emotional valence towards the weather. Based on S's understanding of the literal listener's prior knowledge, she knows that if she produces the utterance "The weather is terrible," $L_{literal}$ will believe that the weather is `terrible` and that the speaker is likely to feel negative about it. Since the QUD is successfully addressed if $L_{literal}$ believes that S feels negative towards the weather state, S is motivated to produce the utterance "The weather is terrible." However, suppose that the pragmatic listener $L_{pragmatic}$ has strong prior belief that the



(a) The nine weather images shown to participants in Experiments 1 and 2.



(b) The proportion of participants who rated each of the weather contexts as each weather state.

Figure 2: Nine weather contexts and their empirically measured priors over weather states.

weather state is `bad`. Since $L_{pragmatic}$ reasons about S and her goals, he realizes that S chose the utterance "The weather is terrible" to communicate her negative affect and not the true state of the weather. He will then infer that the weather is likely `bad`, and that S is extremely likely to feel negative towards it, which successfully produces a hyperbolic interpretation. However, suppose there is strong prior belief that the weather is `amazing`. Given the utterance "The weather is terrible," $L_{pragmatic}$ interprets it literally to mean that the weather is indeed `terrible`. This is because if it were *not* terrible, then it would be unlikely for S to choose the utterance, as it is unlikely to communicate either the true state of the world or her valence The blue lines in Figure 1 show simulations of the model with this one-dimensional affect space. While the model produces a hyperbolic interpretation given strong prior belief about `bad` weather (left panel), it produces a literal interpretation given strong prior belief about `amazing` weather (right panel). In other words, a model that only considers a single affect dimension (valence) is unlikely to infer a positive world state from a highly negative utterance (and vice versa), thus failing on many cases of verbal irony.

We now consider a more complex affect space with two dimensions—valence and arousal—to observe its consequence on interpretation. Affective science identifies valence and arousal as two main dimensions underlying the slew of emotions that people experience (cite). For example, *anger* is a negative valence and high arousal emotion, while *contentment* is a positive valence and low arousal emotion (cite). We suggest that perhaps 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 and 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, valence and arousal. 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`, producing a hyperbolic interpretation. However, given strong prior belief that the weather is `amazing`, the model now interprets "The weather is terrible" ironically to mean 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. These model simulations suggest that reasoning about a two-dimensional affect space motivated by emotion theory enables the model to appropriately interpret ironic utterances.

To produce an interpretation of an utterance in context, the model requires the following input values: (1) $P(s)$: the prior probability of a weather state s given a weather context. (2) $P(a|s)$: the probability of affect a (positive/negative valence and high/low arousal) given a weather state. (3) $P(q)$: the prior probability of a particular QUD (4) The speaker optimality parameter λ. We derived the values for (1) and (2) from Experiment 1 and fit (3) and (4) to the data from Experiment 2, which we describe below.

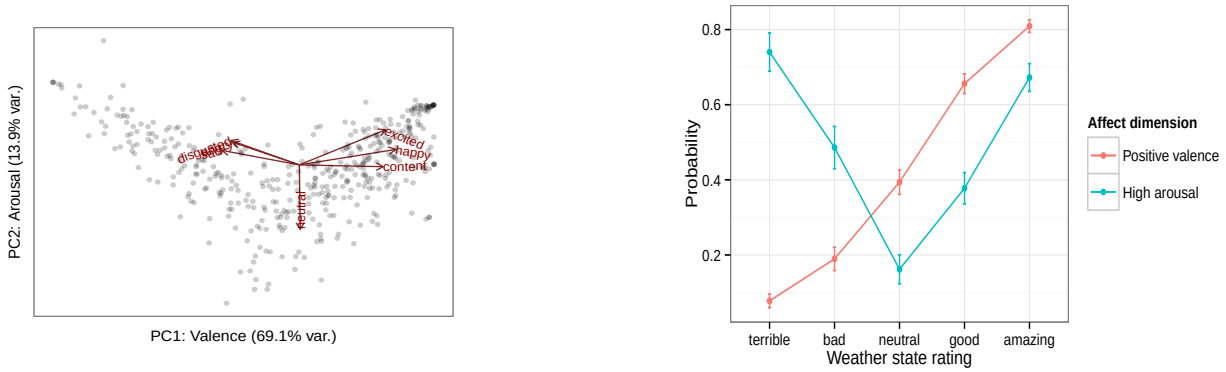
JTK: shou
we includ
these figu

Behavioral Experiments

To quantitatively test our model with the enriched affect space described above, we conducted the following two experiments. In Experiment 1, we measured the prior beliefs over weather states for various weather contexts. We also measured various emotions associated with different weather contexts in order to empirically extract the affective dimensions relevant to 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 (Figure 2a). To cover a range of weather states, three of the images were of sunny



(a) A biplot of the two principal components of the emotion ratings from Experiment 1.

(b) The average probabilities of positive valence and high arousal associated with each weather state. Error bars are 95% confidence intervals.

Figure 3

weather, three of cloudy weather, and three of rainy or snowy weather. We refer to these images as weather contexts.

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*. 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`

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. From Figure 2b, we see 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 affects 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. We found that the first two principal components accounted for 69.14% and 13.86% of the variance in the data (see Figure 3a). In addition, they correspond roughly to the dimensions of emotional valence (positive or negative) and emotional arousal (high or low), respectively, which are consistent with theories of emotion in affective science (cite). To approximate the probabilities of Ann feeling positive or negative affect and high or low arousal given different weather states, we converted the PCA scores into probabilities as follows. We first normalized the scores in each dimension to have zero mean and unit variance. Treating these normalized scores as quantiles of a standard normal distribution, we used the cumulative distribution function to convert the normalized scores into values between 0 and 1. Figure 3b shows the probability of positive valence and high arousal given each weather state.

Experiment 2: Irony understanding

In Experiment 1, we obtained the prior distribution over weather states for each weather context as well as the prior probabilities of positive valence and high arousal given each weather state. This gives us the necessary 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 2a 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`

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 (STATS) (Figure XX?). For example, "The weather is terrible" (a negative utterance) is rated as more ironic in weather context 1 (positive context) than in weather context 7 (negative context). 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-squared 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`, `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 dark blue lines in Figure 4 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 examined 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 predict the projected values of emotion ratings in Experiment 2. We then performed the same transformations on the scores to convert them into values between 0 and 1.

Model Evaluation

In this section, we evaluate the model's performance using results from our behavioral experiments.

(TODO: add actual fit values in a clear manner: $\lambda = 1$, p(state goal) = 0.2, p(valence goal) = 0.3, p(arousal goal)=0.4)

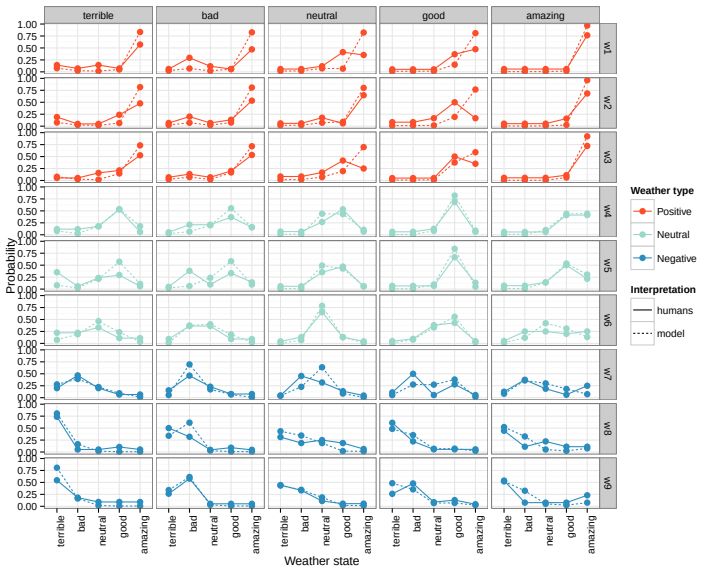


Figure 4: Model's and participants' inferences about the weather state (x-axis) given a weather context (row) and an utterance (column). Each panel represents interpretations of an utterance in a weather context. The solid lines are participants' ratings; the dotted lines are model's posterior distributions over weather states.

The model produced an interpretation for each of the 45 utterance $\times$ weather context pairs. Each interpretation is a joint posterior distribution $P(s,v,a|u)$. We first examine the model's performance on recovering the actual weather state $P(s|u)$ by marginalizing over v and a. Figure 4 shows participants' and the model's interpretation of the actual weather state given an utterance and a weather context. We see that the model predictions closely match humans' interpretations, with a correlation of 0.86 (show scatter plot?). Next, we examine the model's performance on recovering the speaker's valence by marginalizing over s and a. The model's prediction for emotional valence match humans' extremely closely, with a correlation of 0.96. Finally, we examine the model's performance on recovering the speaker's emotional arousal by marginalizing over s and v. The model's prediction for emotional arousal match humans' with a correlation of 0.66 (TODO: add split half; figures???). This suggests that the model is able to incorporate background knowledge and reasoning about multiple affective goals to produce the appropriate ironic interpretations as well as the associated affects.

is the mo
using the
prior too
strongly
here?

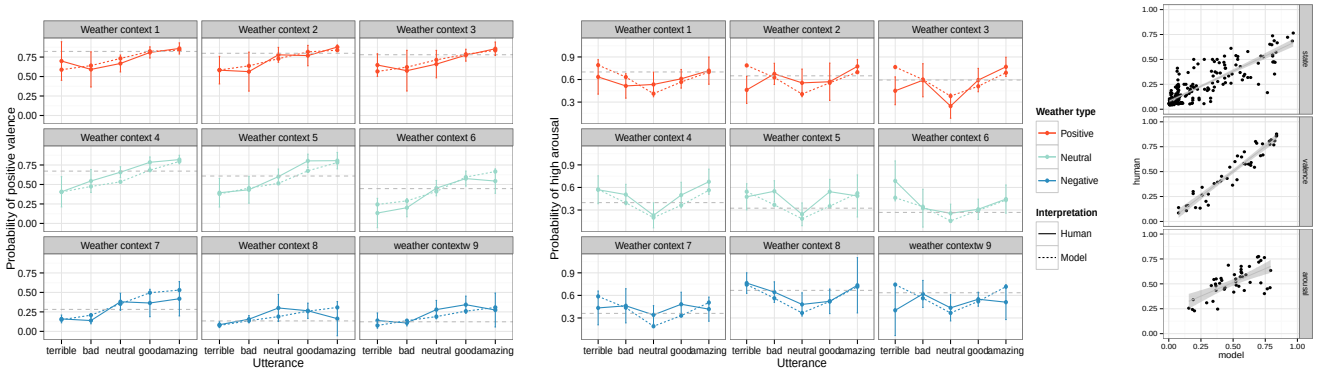
note that
is the cas
even whe
interprete
valence m
matches t
prior ex-
pected va
lence give
literal me
ing – i.e.
model ca
tures iron
valence.

why is thi
correlatio
low? is it
because o

Discussion

summary of take-home point

With our behavioral experiments, we presented results showing the fine-grained effects of background knowledge on people's interpretations of ironic utterances, and identified the



(a) Average probabilities of speaker feeling positive valence given her utterance in a weather context.

(b) Average probabilities of speaker feeling high arousal given her utterance in a weather context.

(c) Scatter plot of human versus model interpretations.

main affective dimensions that are conveyed by irony. In addition, we presented a computational model that predicts peoples' interpretations of ironic utterances using general communicative principles. In effect, the model is able to tell when an utterance is meant to be literal or ironic, and when meant to be ironic, what types of affect the speaker intends to convey. By reasoning about the speaker's communicative goals, the model goes beyond the literal meaning of an utterance to infer the actual state of the world. In addition, it recovers important aspects of the speaker's affect about the situation and captures the social and affective uses of irony. Together, these results suggest that basic principles of communication—background knowledge and reasoning about informativeness with respect to the speaker's communicative goals—may be an important driver of irony understanding.

the above paragraph is murky. the below maybe gets more at our take-away point: the approach in kao2914 extends immediately to additional non-literal uses, by more carefully considering the space of covert meanings that may be conveyed. in general re-work this conclusion to be sharper, and clearer about the main points.

connection to informal theories

Since Grice's original proposal, there have been two competing theories of irony understanding. The echoic mention theory (Sperber & Wilson, 1981; Jorgensen, Miller, & Sperber, 1984) proposes that ironic utterances are intended to remind the listener of a previous utterance that turned out to be false or irrelevant. The affective subtext of irony then arises from the contrast between what was said earlier and what is actually the case. On the other hand, pretense theory (Clark & Gerrig, 1984) 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. Understanding an ironic statement then involves recognizing the pretense as well as the preposterousness of the person being enacted. (could make this whole part shorter)

connection to NLP

Given the prevalence of irony in natural language, many researchers in natural language processing aim to automatically detect sarcasm in large bodies of texts in order to recover the correct sentiment from an ostensibly positive or negative utterance (e.g. "I was overjoyed to pay $30 for an overcooked steak") (Davidov, Tsur, & Rappoport, 2010; Filatova, 2012). A critical insight that emerges from these approaches 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, which computers currently lack. Indeed, without sufficient contextual information, even human beings exhibit poor judgment of verbal irony (González-Ibáñez, Muresan, & Wacholder, 2011; Wallace, Do Kook Choe, & Charniak, 2014).

relationship to hyperbole and what that says about figurative language understanding more generally

It is important to note that the model we present here is only minimally different from the hyperbole model described earlier in the paper (Kao, Wu, et al., 2014). In fact, the only difference is that instead of considering a single affect dimension (negative valence), this extended model takes seriously the circumplex model of affect (cite), which identifies valence and arousal as two main dimensions underlying the slew of emotions people experience. The similarities between these two models suggests that hyperbole and irony may be understood using similar principles of communication. A deeper point we wish to make with this work is that communication in general, and nonliteral language understanding in particular, relies on reasoning about the speaker's communicative goals during interpretation. Furthermore, these goals are often social or affective in nature, and speakers are adept at harnessing shared background knowledge with the listener to convey rich affective meanings without explicitly stating them.

future directions and conclusion

Our experimental paradigm and modeling framework lends itself well to a more detailed and precise account of irony understanding. For example, what are the range of social and af-

fective meanings involved in irony understanding? We identified emotional valence and arousal in these experiments, but there may be others ....(attitudes? opinions? affect? social closeness?) (TODO: finish) What are the social functions of irony? Irony is often used to signal social closeness with the listener, presumably because it expresses the speaker's assumption that she and the listener share a great deal of common ground (cite). Given an ambiguous utterance that could be interpreted either literally or ironically, if a speaker supplements the utterance with prosodic cues to signal that it is meant ironically, will listeners judge the speaker as being closer to them? Can our model more directly test the predictions of the echoic mention and pretense theories of irony and help distinguish between them? Given the prevalence of irony in everyday language, we believe it would be *amazing* (literally) to address these questions in future research and understand irony's mysterious ability to be interpreted as the opposite of what it is.

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).

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.

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

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.