The Pragmatics of Metaphor Understanding: A Computational Approach

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

Leon Bergen (bergen@mit.edu)
Department of Brain and Cognitive Sciences
MIT, USA

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

Abstract

Abstract goes here.

Keywords: language understanding; metaphor; pragmatics; computational models

Introduction

Nonliteral language is, quite literally, everywhere. Human communication is laden with metaphor, irony, and hyperbole, often creating poetic or humorous effects that add rich and important dimensions to language (Glucksberg, 2001; Pilkington, 2000; Lakoff & Turner, 2009; Roberts & Kreuz, 1994; B. Bergen & Binsted, 2003). Of the various types of nonliteral language, metaphor has inspired a particularly abundant amount of research in cognitive science, ranging from how metaphors structure and shape our thoughts (Ortony, 1993; Lakoff et al., 1993; Thibodeau & Boroditsky, 2011), whether metaphor processing recruits the same strategies as standard language processing (Giora, 1997; Ortony, Schallert, Reynolds, & Antos, 1978; Gibbs Jr, 2002; Glucksberg & Keysar, 1993) and what factors determine the meaning and aptness of a novel metaphor (Blasko & Connine, 1993; Tourangeau & Sternberg, 1981; Kintsch & Bowles, 2002). The overwhelming amount of interest in metaphor research is due in part to the ubiquity of metaphor in everyday language as well as the belief that metaphor may be critical for helping us understand how the mind creates meaning (Ricoeur, 2003; Gibbs, 1994).

One area of research focuses on the pragmatic principals that listeners utilize to infer meaning from metaphorical utterances (Tendahl & Gibbs Jr, 2008; Stern, 2000). Rather than view metaphor as a separate mode of communication that requires specialized language processing strategies, this approach argues that basic principles of communication drive the meaning that a listener infers from a metaphor (Sperber & Wilson, 2008). Relevance theory, in particular, posits that listeners interpret utterances with the assumption that speakers produced them because they are maximally relevant. Relevance theorists argue that this principle explains how listeners infer the meaning of a novel metaphor as well as other forms of loose talk where the meaning of the utterance is underspecified (Wilson & Sperber, 2002; Wilson & Carston, 2006; Sperber & Wilson, 1985). When interpreting the metaphor "My lawyer is a shark," for example, the listener assumes that the speaker aims to communicate features of "a shark" that are relevant to the person under discussion ("my lawyer"), and thus do not access features such as *has fins* or *swims*.

While many linguists and psychologists have argued for the benefits of studying metaphor using a pragmatics framework, to our knowledge there is no formal model showing that effects in metaphor understanding may arise from basic principles of communication. On the other hand, a recent body of work presents a series of computational models for pragmatic reasoning, where speaker and listener recursively reason about each other to communicate effectively (Frank & Goodman, 2012; Jäger & Ebert, 2009). By formalizing principals of communication, these Rational Speech Act models are able to quantitatively explain a range of phenomena in language understanding, such as scalar implicature and the effect of alternative utterances (Goodman & Stuhlmüller, 2013; L. Bergen, Goodman, & Levy, 2012). However, a limitation of these models is that they are unable to predict interpretations of an utterance that are false under its literal meaning. In this paper, we extend the model to consider multiple dimensions of meaning. The listener assumes that the speaker chooses an utterance to maximize informativeness along dimensions of meaning that are relevant to the conversation. This makes it possible for a literally false utterance to be optimal as long as it is informative along the target dimension. Critically, this framework closely aligns with the relevance-theoretic view that a listener considers the relevance of a meaning to the question under discussion in order to infer what the speaker intended to communicate.

Although metaphor understanding is a complex phenomenon that calls for a variety of approaches, we argue that the interpretation of at least some types of metaphor are shaped at least in part by basic principals of pragmatics. Building upon Rational Speech Act models, we present a computational model showing that rich metaphorical meaning can arise from basic pragmatic reasoning.

To reasonably limit the scope of our work, we focus on metaphors of the classic form "X is a Y." We describe a computational model that can interpret such sentences metaphorically and conduct behavioral experiments to evaluate the model's performance. We show that a listener's interpretation of a metaphor is driven by context and the question under discussion, and that this effect is captured by our formalization of the relevance principal. Finally, we also show that metaphors often communicate information more efficiently than literal statements and hence can be optimal and rational speech acts. [Or whatever we end up focusing on in the error analysis]

Computational Model

At the core of basic Rational Speech Act models, a listener and a speaker recursively reason about each other to arrive at pragmatically enriched meanings. Given an intended meaning, a speaker reasons about a literal listener and chooses an utterance based on its informativeness. A pragmatic listener then reasons about the speaker and uses Bayes Rule to infer the meaning given the utterance.

We extend this model and formalize a notion of relevance by considering the idea that a speaker may have different communicative goals. Intuitively, an utterance is optimally informative and relevant if it satisfies the speaker's communicative goal. Since the speaker's precise communicative goal may be unknown to the listener, the listener performs joint inference on the goal as well as the intended meaning.

By introducing multiple potential goals for communication, we open up the possibility for a speaker to produce an utterance that is literally false but still satisfies her goal. Importantly, we argue that the speaker achieves this by exploiting her and the listener's prior knowledge–their common ground–to reason about what information the listener would gain if he takes the utterance literally. To illustrate this idea more concretely and demonstrate how it is implemented in our model, we will use the metaphor "John is a shark" as an example. Suppose the speaker's goal is to communicate that John has the feature *fierce*. She reasons about a literal listener who will interpret an utterance "John is a shark" as meaning that John is literally a member of the category "shark." Based on the speaker's understanding of the literal listener's prior knowledge, she knows that the literal listener will very likely believe that John, who he believes is literally a shark, is fierce. If the listener believes that John is fierce, then the speaker's goal is satisfied. As a result, the speaker is motivated to produce that utterance in order to optimize informativeness and relevance. A pragmatic listener now reasons about such a speaker. Based on prior knowledge, he knows that John is extremely unlikely to be literally a member of the shark category. On the other hand, he knows that the speaker is fairly likely to want to communicate about John's fierceness. He also knows that the speaker knows that *fierce* is a high-probability feature of sharks. Given his prior knowledge, his model of the speaker, and the utterance she produces, he infers that the meaning of the utterance is likely to be that John is a fierce person.

We now formalize this intuition mathematically. For simplicity, in this model we restrict the number of possible categories to which a member may belong to c_a and c_p, denoting an animal category or a person category, respectively. We also restrict the possible features under consideration to a vector of size three: $\vec{f} = \{f_1, f_2, f_3\}$, where f_i is either 0 or 1. If the speaker has a goal g_i, then she wishes to maximize informativeness about f_i.

The literal listener is modeled as

$$L_0(c, \vec{f}|u) = \begin{cases} P(\vec{f}|c) & \text{if } c = u \\ 0 & \text{otherwise} \end{cases}$$

The speaker is modeled as

$$S_n(u|g) \propto e^{U_n(u|g)}$$

Optimizing the probability of the speaker's goal being satisfied can be accomplished by minimizing the goal's information-theoretic surprisal given an utterance. Given an utterance u, the listener L_n will guess that the meaning is $c, \vec{f}$ with probability $L_n(c, \vec{f}|u)$. The probability of the speaker's goal being satisfied is therefore the following:

$$\sum_{c,\vec{f}} L_n(c, \vec{f}|u) g(\vec{f})$$

Since the utility function U_n is composed of both the negative surprisal of the goal and the negative of the utterance cost, combined with equation 2 and using a uniform utterance cost, this leads to:

$$S_n(u|g) \propto \sum_{c,\vec{f}} L_n(c, \vec{f}|u) g(\vec{f})$$

The listener L_n performs Bayesian inference to guess the intended meaning given the prior knowledge and his internal model of the speaker. To determine the speaker's intended meaning, the listener will marginalize over the possible goals under consideration.

$$L_n(c, \vec{f}|u) \propto \sum_{g} P_C(c) P_F(\vec{f}|c) P_G(g|c, \vec{f}) S_{n-1}(u|g)$$

Note that the listener needs to consider the following prior probabilities:

(1) $P_C(c)$, the prior probability that X belongings to category c. We assume that X belongs to c_p with very high probability ($P_C(c_p) = 0.9999$).

(2) $P_F(\vec{f}|c)$, the prior probability that a member of category c has feature values $\vec{f}$. We obtain this empirically in Experiment 1.

(3) $P_G(g|\vec{f})$, the prior probability that given that a speaker knows the value of the feature vector $\vec{f}$, she wishes to communicate goal g. We assume that this prior can change given the question under discussion, i.e. the context that a question sets up.

Behavioral Experiments

To obtain human metaphorical interpretations that we can compare against our model, we focused on a set of 32 comparing human males to different non-human animals. We selected 32 common non-human animal categories from English Club (url). Using this list, we conducted Experiment 1A to elicit a set of three salient features for each animal category. We conducted Experiment 1B to elicit the feature priors $P_F(f|c)$ described in the model section (see Table 1). Finally, we conducted Experiment 2 to measure people's interpretations for the set of metaphors.

Experiment 1A: Feature Elicitation

Materials and Methods 100 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each subject read 32 animal category names presented in random order, e.g. "whale", "ant", "sheep". For each animal category, subjects were asked to type the first adjective that came to mind in a text box.

Results Using subjects' responses, we constructed a list of adjectives for each animal category and ordered them by the number of times they were given by a different subject (i.e. their popularity). We removed all color adjectives, such as "brown" and "black." To avoid constructing a set of features that have roughly equivalent meanings such as "big", "huge", and "large", we used Wordnet to identify synonymous adjectives and only kept the most popular adjective among a set of synonyms. We then took the top three most popular adjectives for each animal category and used them as the set of features. Note that $f1$ is the most popular adjective, $f2$ the second, and $f3$ the third. Table 1 shows the animal categories and their respective features.

Experiment 1B: Feature Prior Elicitation

Materials and Methods We used Wordnet to construct antonyms for each of the adjective features produced in Experiment 1A. When multiple antonyms existed or when no antonym could be found on Wordnet, the first author used her judgment to choose the appropriate antonym. Table 1 shows the resulting list of antonyms. For each animal category, eight possible feature combinations were constructed from the three features and their antonyms. For example, the possible feature combinations for a member of the category "ant" are {small, strong, busy}, {small, strong, idle}, {small, weak, busy}, and so on.

60 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each subject completed 16 trials in random order. Each trial consisted of the eight feature combinations for a particular animal category. Using slider bars with ends marked by "Impossible" and "Absolutely certain," subjects were asked to rate how likely it is for a member of the animal category to have each of the eight feature combinations. Subjects also rated the probabilities of the feature combinations for a male person.

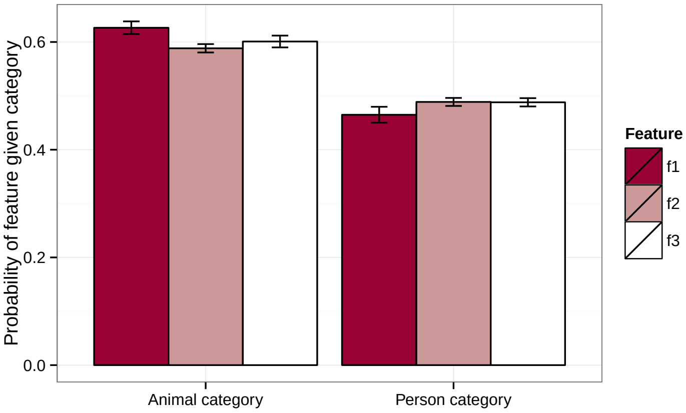

Figure 1: This is a figure.

Results We make the simplifying assumption that the eight feature combinations presented in each trial exhaustively describe a member of a particular category. As a result, we normalized each subject's ratings for the eight feature combinations in a trial to sum up to 1. Averaging across subjects' normalized ratings, we obtained the feature priors $P_F(\vec{f}|c)$ for $c = c_a$ (animal) and $c = c_p$ (person), assuming that $f_i = 1$ is represented by the feature adjective and $f_i = 0$ is represented by the antonym.

For ease of interpretation, in Table 1 we present the marginal probabilities of each of the three features instead of the joint probabilities. Figure 1 shows the average marginal probabilities of features given an animal category versus a person category. We see that by design, features are rated as significantly more likely to be present given the animal category than the person category.

Experiment 2: Metaphor Understanding

Materials and Methods We created 32 scenarios based on the animal categories and results from Experiment 1. In each scenario, a person (e.g. Bob) is having a conversation with his friend about a person that he recently met. Since we are interested in how relevance to the question under discussion (QUD) affects metaphor interpretation as well as the effectiveness of metaphorical versus literal utterances, we created four conditions for each scenario by crossing vague/specific QUD and literal/metaphorical statements. In vague QUD conditions, Bob's friend asks a vague question about the person Bob recently met: "What is he like?" In specific QUD conditions, Bob's friend asks a specific question about the person: "Is he $f1$?" Where $f1$ is the most popular adjective for a given animal category c_a in Experiment 1A. In literal conditions, Bob replies with a literal utterance,, either by saying "He is $f1$" to the question "What is he like?" or "Yes" to the question "Is he $f1$?". In Metaphorical conditions, Bob replies with a metaphorical statement, e.g. "He is a c_a" where c_a is an animal category. See Table 2 for examples.

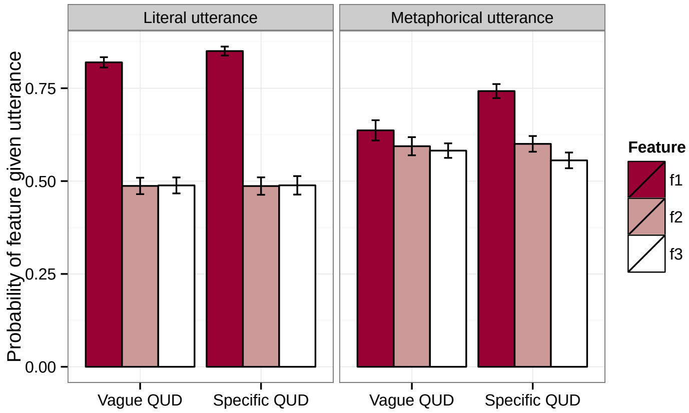

Figure 2: This is a figure.

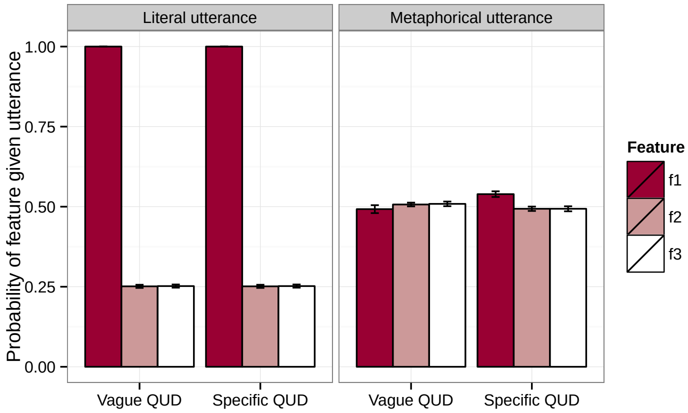

Figure 3: This is a figure.

Results 49 native English speakers with IP addresses in the United States were recruited on Amazon's Mechanical Turk. Each subject completed 32 trials in random order. The 32 trials were randomly and evenly assigned to one of the four conditions, i.e. each subject read 8 scenarios for each condition. For each trial, subjects used sliders to indicate the probabilities that the person described has features $f1$, $f2$, and $f3$.

For each condition of each scenario, we obtained the average probability ratings for the three features. Figure 2 shows the average ratings for each feature across animal categories given a vague or specific QUD and a literal or metaphorical utterance. We see that when the speaker gives a literal statement directly affirming the presence of $f1$, subjects rate $f1$ as significantly more likely than when the speaker gives a metaphorical statement. However, subjects rate $f2$ and $f3$ as significantly more likely when the speaker produces a metaphorical utterance. We also see an effect of the QUD on the interpretation of metaphorical utterances. Given a specific question about $f1$, subjects interpret the speaker's metaphorical utterance as being relevant to the question and rates the probability of $f1$ as significantly higher than when the QUD is vague. On the other hand, the probabilities of $f2$ and $f3$ are not significantly different given a vague QUD or a specific QUD about $f1$.

Model Comparison

We used the feature priors obtained in Experiment 1B to compute model interpretations of the 32 metaphors. For the category prior $P_C(c)$, we assumed that given the common ground set up by the conversation, it is extremely unlikely for the person described to actually belong to the animal category ($P_C(c_a) = 0.0001$) and extremely likely for him to belong to the person category ($P_C(c_p) = 0.9999$). We model the effect of relevance to the question under discussion by assuming that the goal prior $P_G(g|\vec{f})$ varies given vague or specific QUDs. When the QUD is vague, we set the distribution as uniform over goals that are consistent with $\vec{f}$. When the QUD specifically addresses f_1, we set the distribution as having a much higher probability for g_1 and equal probability for g_2 and g_3.

Using these prior settings and the model we described, we obtained feature probabilities for each of the 32 metaphors. Figure 3 shows the average marginal feature probabilities for the 32 metaphors given a vague or specific QUD. We see that the model captures the QUD effect, where $f1$ receives a significantly higher probability when the speaker is a priori more likely to be informative about $f1$. (Need to describe results for "literal" statement where it's just the prior.)

To quantitatively evaluate the model's performance on metaphorical utterances, we correlated model predictions with human ratings for each of the features given a metaphorical utterance and a vague or specific QUD. We first focused on the model's performance on $f1$ features, namely the most salient features and ones that can be specifically under discussion. Correlation between human ratings and the model's marginal posterior probabilities for $f1$ across the 32 metaphors and vague/specific QUD conditions is 0.73 (add Spearman prophecy formula), suggesting that the model captures a significant amount of the reliable variance in the human data. We then compare this performance with baseline models that only consider feature priors of the source (animal category) or target (person category). A baseline model consisting of only feature priors for the animal categories yields a non-significant correlation ($r = -0.03, p > 0.05$). A baseline model consisting of only feature priors for the person category yields a significant correlation ($r = 0.59, p < 0.01$), but one that is significantly worse than the model predictions ($p < 0.001$ with a Cox test). A linear regression model that takes both sets of priors as predictors still yields a significantly worse fit than our model. This suggests that our model adequately combines prior knowledge about the source and target domains to produce metaphor interpretations that closely fit humans'.

We now evaluate the model's performance on all three

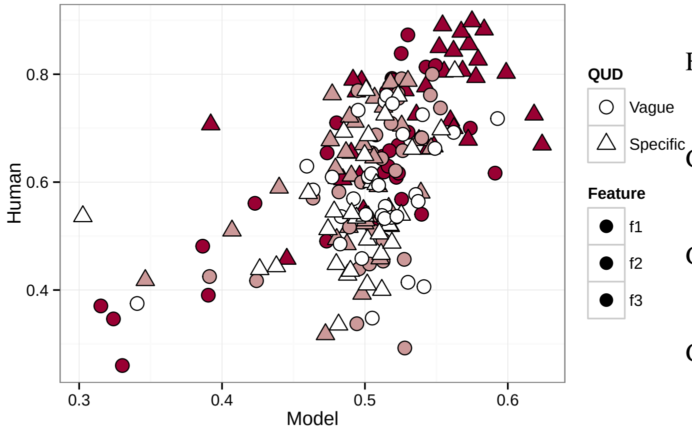

Figure 4: This is a figure.

types of features. Correlation between human ratings and the model's marginal posterior probabilities for $f1$, $f2$, and $f3$ across the 32 metaphors and vague/specific QUD conditions is 0.56 (add Spearman prophecy formula). Figure 4 shows the model predictions against human ratings for the three features of each metaphor given a vague or specific QUD. While our model still captures a significant amount of reliable variance in the human data, we see that there are certain features, particularly $f2$ and $f3$, for which the model performance is significantly worse. We analyze these metaphors and features in more detail in the following section.

Discussion

Discuss implication of results on the pragmatics of metaphor; discuss other effects we could explore using the modeling framework; suggest future directions.

In this paper we focus on developing a computational model of pragmatics that explains a range of effects in metaphor understanding, with the goal of advancing our understanding of the computational basis of metaphor and nonliteral language understanding.

Footnotes

Acknowledgments

Place acknowledgments (including funding information) in a section at the end of the paper.

References

Bergen, B., & Binsted, K. (2003). The cognitive linguistics of scalar humor. *Language, culture, and mind*, 79–92.

Bergen, L., Goodman, N. D., & Levy, R. (2012). Thats what she (could have) said: How alternative utterances affect language use. In *Proceedings of the thirty-fourth annual conference of the cognitive science society.*

Blasko, D. G., & Connine, C. M. (1993). Effects of familiarity and aptness on metaphor processing. *Journal of experimental psychology: Learning, memory, and cognition*, *19*(2), 295.

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

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.

Gibbs Jr, R. W. (2002). A new look at literal meaning in understanding what is said and implicated. *Journal of Pragmatics*, *34*(4), 457–486.

Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, *8*, 183–206.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms.* Oxford University Press.

Glucksberg, S., & Keysar, B. (1993). How metaphors work.

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

Jäger, G., & Ebert, C. (2009). Pragmatic rationalizability. In *Proceedings of sinn und bedeutung* (Vol. 13, pp. 1–15).

Kintsch, W., & Bowles, A. R. (2002). Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, *17*(4), 249–262.

Lakoff, G., et al. (1993). The contemporary theory of metaphor. *Metaphor and thought*, *2*, 202–251.

Lakoff, G., & Turner, M. (2009). *More than cool reason: A field guide to poetic metaphor*. University of Chicago Press.

Ortony, A. (1993). *Metaphor and thought*. Cambridge University Press.

Ortony, A., Schallert, D. L., Reynolds, R. E., & Antos, S. J. (1978). Interpreting metaphors and idioms: Some effects of context on comprehension. *Journal of Verbal Learning and Verbal Behavior*, *17*(4), 465–477.

Pilkington, A. (2000). *Poetic effects: A revelance theory perspective* (Vol. 75). John Benjamins.

Ricoeur, P. (2003). *The rule of metaphor: The creation of meaning in language*. Psychology Press.

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

Sperber, D., & Wilson, D. (1985). Loose talk. In *Proceedings of the aristotelian society* (Vol. 86, pp. 153–171).

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. *The Cambridge handbook of metaphor and thought*, 84–105.

Stern, J. J. (2000). *Metaphor in context*. The MIT Press.

Tendahl, M., & Gibbs Jr, R. W. (2008). Complementary perspectives on metaphor: Cognitive linguistics and relevance theory. *Journal of Pragmatics*, *40*(11), 1823–1864.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS One*, *6*(2), e16782.

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive psychology*, *13*(1), 27–55.

Wilson, D., & Carston, R. (2006). Metaphor, relevance and the emergent propertyissue. *Mind & Language*, *21*(3), 404–433.

Wilson, D., & Sperber, D. (2002). Relevance theory. *Handbook of pragmatics*.