Loyal as a Fox:
Formalizing the Role of Typicality, Salience, and Alternative Utterances in Metaphor Understanding

Alexandra Mayn and Vera Demberg

Department of Computational Linguistics and Phonetics,
Saarland University, 66123 Saarbrücken, Germany
s8almayn@stud.uni-saarland.de, vera@coli.uni-saarland.de

Abstract. We propose a computational model of metaphor understanding within the Rational Speech Act framework which views metaphor interpretation through the lens of gradient salience. The model predicts that other features of a referent, as well as possible alternative referents have an impact on how a metaphor is interpreted; that it can be interpreted inversely when the feature in question is saliently atypical for it, and, finally, that when a metaphor is used to convey a certain feature and it has an average typicality for that feature, that will make it confusing and prone to misinterpretation. We run our model on a small demo dataset and propose a set of behavioral experiments to evaluate the model's predictions against human judgment.

Keywords: Metaphor Interpretation · Rational Speech Act Theory · Salience.

1 Introduction

Metaphors abound in natural language and are therefore an important phenomenon to capture when modeling processing of discourse. There is a growing body of literature investigating interpretation, processing, and difficulty of metaphors[1]. There have also been attempts to model metaphors probabilistically. Kao et al.[2] were the first to formalize metaphor understanding using the Rational Speech Act Framework[3], which has been successful at modeling a variety of nonliteral language uses in context, including politeness[4] and puns[5].

Kao et al.[2] limit their scope to metaphors of type "X is a Y", where X is a human and Y is a member of an animal category. The crux of their idea is that the speaker wishes to communicate features which are characteristic of this animal and are relevant to humans. For instance, if I tell you that "John is a shark" and you know who I am referring to, you are unlikely to think that John actually has fins and lives in the ocean. Instead, you will probably think that I want to communicate a feature of John's personality by invoking an animal of whom this feature is representative; in this case, probably scariness or meanness.

We propose a novel model of metaphor understanding which builds on Kao et al.'s work but takes a graded approach and represents features of the metaphor in terms of their salience[1]. We also focus on metaphors of the form "X is a Y", where X is a human male and Y is an animal category. Kao et al. define each category by a vector of 3 binary features which are different for every animal category. Our model, on the other hand, defines a member of a category as having all possible features to some degree (a real value between 0 and 1), depending on how salient a given feature is for that animal. This gradient approach captures several important aspects of metaphor interpretation: it takes into account alternative interpretations of an utterance by capturing relative salience of features within an animal ("an ant is more saliently strong than it is fast") as well as alternative utterances for the same intended meaning ("an ox is more saliently strong than an ant"). We derive salience from typicality with the assumption that both very high and very low typicality are salient, whereas average typicality is not and is therefore less likely to be referred to. This presupposes, so to speak, "negative" or inverse use of metaphors which Kao et al. do not consider, e.g. saying "John is a fox" to mean that John is *not* loyal because loyalty is very atypical (and therefore salient) for a fox.

We demonstrate the model's predictions on a small demo dataset. In order to compare the model's predictions to human judgments, we are in the process of conducting behavioral experiments to elicit priors and their interpretations in context.

The rest of this paper is structured as follows. Section 2 summarizes the Rational Speech Act framework of which our model is an instance, and describes the proposed model. Section 3 outlines behavioral experiments that will be used to evaluate the model. Section 4 demonstrates the model's predictions on a small demo dataset (available in Section 6) and includes a sketch of planned comparisons to human performance and Kao et al.'s model. Section 5 includes the discussion of the model, concludes, and suggests avenues for future work.

2 Computational Model

We model metaphor understanding using the Rational Speech Act framework introduced by Frank & Goodman[3] which has been successful at modeling a variety of nonliteral language uses in context including politeness[4], puns[5], and metaphors[2]. At its core is the idea of the listener and the speaker recursively reasoning about each other to arrive at a common interpretation: a first-level pragmatic speaker reasons about a literal listener and acts in such a way as to maximize her utility, which is based on the informativeness of the utterance. In turn, the pragmatic listener reasons about the pragmatic speaker to recover the intended utterance. Kao et al.[2], who first formalized metaphor understanding using RSA, extended the model to include communicative goals in order to incorporate literally false utterances into the model, which we also adopt. We limit the scope of the types of metaphors to the type X is a Y, where X is a male

[1] The model along with the dataset used for this paper is available at `https://github.com/sashamayn/rsa_metaphors`

name and Y is an animal category; in this way our methodology and findings will form a natural extension to Kao et al.s work.

Kao et al. define possible features of X as a vector of features of length 3, in which each of the features are binary. The limitations of that approach are that, first, it does not represent relative salience of features both between and within animals (i.e., of types an ant is more saliently strong than it is fast and an ox is more saliently strong than an ant), which means that it does not take into account alternative utterances and alternative interpretations of a given utterance, which, we suppose, both play a crucial role in metaphor interpretation.

Therefore, we define an animal as a vector of size [total number of features], where each of the features is its normalized salience for a given animal. This approach also allows for negative use of metaphors, e.g. interpreting "He is a fox" as "He is **not** loyal", since foxes are associated with being cunning and therefore have a low typicality (and, therefore, high salience) value for loyalty.

Based on the above considerations, we define the literal listener L0 as:

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

where f is the feature in question, c is the animal, and u is the utterance. The literal listener will hear the utterance "John is a shark" and interpret it as John literally belonging to the category "shark" and having the feature in question to the extent to which it is salient for a shark. We are not the first work to define the literal listener gradiently. Graf et al.[6] define their literal listener as the degree of acceptability of an object for a given category.

The pragmatic speaker acts in accordance with her utility, which is defined as the log of the literal listener. This is similar to the literal listener equation in Kao et al. but without a projection of the feature vector along one feature dimension since only one feature is being considered:

$$U(u|f) = log(L_0)$$

The choice of the utterance is governed by a softmax choice rule based on utility of the utterance given the feature that the speaker wants to communicate:

$$S_1(u|f) \propto e^{\lambda U(u|f)} = L_0^{\lambda}$$

where λ is an optimality parameter that is fit to the data. As Kao et al. point out, since the speaker's goal is to communicate the strength of the feature (the scariness of a shark), S_1's goal is satisfied because she knows that the literal listener L_0 will interpret the utterance as the person being literally a shark and therefore having the feature of being scary.

Finally, we define the pragmatic listener as:

$$L_1(c, f|i, u) \propto S_1(u|f) \cdot L_1(f, i|c) \cdot L_1(c)$$

The derivation of the equation is included in Section 8. The definition of the pragmatic listener consists of the following components:

- $S_1(u|f)$ the speaker's probability of uttering a certain animal category given her intention
- $L_1(c)$: the prior probability that the entity in question is either a human or a non-human animal. Following Kao et al., we set $L_1(c = c_a) = 0.01$ for every animal and $L_1(c = c_h) = 0.99$. We assume that the only two interpretations the listener is considering are a person and that animal.
- $P(f, i|c)$: A joint probability of the speaker wanting to communicate feature f and the conversational context i - speaker's intention as understood by the listener - given that the referent belongs to the category c.

  The two conversational contexts we consider are *specific*, when the listener assumes that the speaker seeks to communicate a specific feature, and *vague*, when it is unclear to the listener which feature the speaker is trying to communicate. We obtain it by multiplying the $salience(f, c)$ of feature f for that category by a factor n: we want to boost the probability $P(f, i|c)$ when $i = f$ (the listener inquiring about a certain feature will make the speaker more likely to want to communicate it), and have it be small but nonzero if $i \neq f$. For the demo dataset, we set $n_{match} = 4$ for the case where $f = i$ and $n_{mismatch} = 0.25$ for the case where $f \neq i$, but these parameters can be fit to the data. We do not scale the case where the conversational context is vague. We then normalize the resulting values to obtain a probability distribution.

2.1 Deriving Salience from Typicality

Frank and Goodman[3] define salience as the prior probability that an object will be referred to in discourse. We make the assumption here that when a feature is a spectrum, both ends of that spectrum are salient, and, therefore, an animal is equally likely to be referred to if the feature in question is very *atypical* of it. This needs to be confirmed experimentally and then the hyperparameter μ can be fit to the data.

We define the salience of a feature for an object as follows:

$$salience(f, o) = |typicality(f, o) - \mu|^{\kappa}$$

where (normalized) typicality is a real number between 0 and 1, and μ and κ are hyperparameters to be fit to the data. μ shifts the typicality values to be centered around 1-μ; κ allows for a better fit of the function to the data. The salience matrix is then normalized by column (animal), so that salience of a feature can be interpreted as a probability that it will be referred to.

For the demo dataset, we set $\mu = 0.5$ and $\kappa = 4$. It is noteworthy that here we are making a tentative assumption that salience is symmetric for the two ends of the typicality spectrum. However, it does not have to be the case. In a different context, Frank[7] found that people did not violate literal semantics to refer to an object when doing so was the only way to uniquely identify it; that is, they did not utter X to mean "the object that does not have an X". Therefore, it is possible that salience is asymmetric and that high atypicality of a feature is less salient than typicality that is at an equal distance from the mean. This

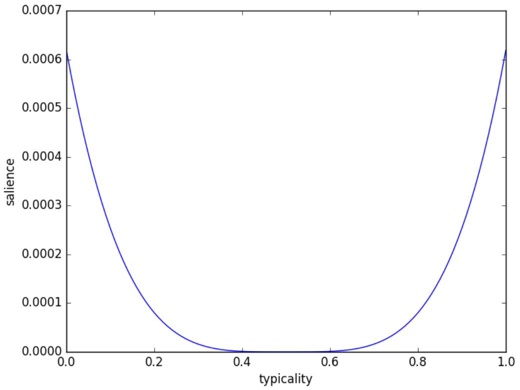

Fig. 1. Salience for a given typicality value with $\mu = 0.5$ and $\kappa = 4$

will be investigated in the behavioral experiments and then μ can be fit to the data accordingly.

3 Behavioral Experiments

We propose a series of behavioral experiments to elicit people's interpretation of metaphors in the vague and specific goal conditions with varying typicality.

3.1 Free-Response Feature Elicitation

MATERIALS A preliminary list of 20 adjectives describing human personality traits. The adjectives were selected from Kao et al.'s longer list where not all features were applicable to humans, and 4 additional ones were included. Out of that list of 20 adjectives, two lists were created containing 10 of the original adjectives and opposites of the remaining 10 (e.g., "disloyal" for "loyal".)
METHOD 10 native English speakers read each of the two lists of 20 characteristics and were asked to type in whatever animals they associate with the trait in question. They were required to fill in at least 10 of the 20 fields. That was done in order to determine which adjectives collected more responses and were therefore useful to keep.
RESULTS Based on the participant responses, the list was be edited in the following way: if the participants' responses to an adjective had high entropy (many responses with little overlap), the adjective was excluded. Similarly, the authors used their judgment to draw the cut-off line for the animals which will be included in the future experiments. If both opposites (e.g., "loyal"-"disloyal") had low entropy, the one with respectively higher entropy was removed from the list.

3.2 Closed-Set Feature Elicitation

MATERIALS The lists of animals and adjectives obtained in the free-response feature elicitation.
METHOD 20 native English speakers were presented with the adjectives one at a time and a list of animals (from 3.1) and were asked to click on the animals they associate with the adjective. Additionally, a text box was available to type in animals that are not on the list if the participants felt like an animal of whom the feature was typical was missing form the list.
RESULTS Based on the participant responses, the lists will be further adjusted by excluding animals the participants rarely used or adjectives with high entropy of responses.

3.3 Typicality Elicitation

MATERIALS Using the feature list obtained in the experiment above, at this stage we will elicit typicality ratings from a separate group of 20 native English speakers.
METHOD Each participant will see every animal category along with the full list of features. They will then use a 7-point Likert scale to answer the question: ”How typical is this feature for this animal?”, 0 being ”not at all typical” and 6 being ”highly typical”. Typicality of each of these features for human males will also be elicited. We use only male names to be consistent with Kao et al.
RESULTS The typicality ratings for each feature for each animal category (e.g., friendliness for a dolphin) will then be averaged and normalized (divided by 6, the maximum typicality value), resulting in a typicality rating for each animal category, including human.

3.4 Metaphor Interpretation

MATERIALS Based on the animal categories and typicality ratings obtained in the previous experiments, we now proceed to the main experimental question. Since we are interested in the role of typicality and salience, as well as of communicative goal, on metaphor understanding, for each non-human animal category we we will create 4 scenarios (1 for vague and 3 for the specific communicative goal) of the type used by Kao et al.[2], in which Bob is talking to his friend about a person he recently met.

In the case of vague communicative goal, Bob's friend asks the vague question ”What is he like?”, to which Bob replies by saying ”He is a c_a”. For the specific goal condition, the friend asks a question mentioning a specific feature, ”Is he f_i*?*”. There are three possible scenarios in this condition - an animal category for which the feature in question is extremely typical, moderately typical, and extremely atypical. What the two extreme cases share and what the averagely typical case lacks is high salience. Table 1 includes examples of each condition.
METHOD 24 native English speakers who did not take part in any of the previous three experiments will participate in this experiment. Every participant

Table 1. Utterance conditions

Goal	Typicality	Example question	Example utterance
vague		”What is John like?”	”He is an ox.”
specific	high	”Is John loyal?”	”He is a dog.”
specific	average	”Is John loyal?”	”He is a dolphin.”
specific	low	”Is John loyal?”	”He is a fox.”

will see all data points in the vague condition and one data point for each animal in a specific condition (20 x 2 = 40 trials in total). This will result in 24 data points per animal in the vague condition and 8 data points for each combination of animal and feature with a given typicality level (e.g., dog-loyal).

For each trial, the participants will see a question-answer pair and will be asked to interpret the metaphor. First, they will select whether they think John is an animal or a person. Second, they will be asked: ”What did Bob mean to say about John?”. They will have to select from a drop-down list of features, and for each feature they select, they will use a 7-point Likert scale to indicate the degree to which the feature is present in John (e.g. ’0 - John is not at all loyal’ to ’6 - John is very loyal’.) Finally, for each data point they will answer the question: ”How certain are you that that is what Bob meant?” by reporting a number on a 7-point Likert scale.

RESULTS For the vague goal condition and for each typicality subcondition of the specific goal we obtain the following characteristics (by averaging):

- The percent of responses that John is a person.
- The features interpreted and the degree to which they are present in John.
- Average certainty ratings (out of 6, later normalized).

We then perform the following statistical comparisons to test the model’s predictions:

1. **Certainty that John is a human:** Proportion of responses that John is a human. We expect that to be very high and not significantly different between the 4 subconditions. That would illustrate that people were able to deduce that the utterance was being used metaphorically.
2. **Informativeness:** The way in which the utterance informed us about John is represented by the difference between the priors for humans and the ratings of the degree to which the feature is present in John. We expect, first of all, the averages for the vague condition and for the specific high typicality subcondition to be positive (the listener learned that the feature is present in John) and the average for the low typicality condition to be negative (the listener learned that the feature is absent in John, i.e. that John is **not** loyal). It is again harder to make predictions about the average typicality case; it is probable that the difference is positive based on our tentative hypothesis that people use metaphors ”positively” more often. In terms of the absolute value of the change, we expect it to be significantly smaller for

the average typicality condition (it is more confusing and less informative) than for either of the other two specific subconditions and for the vague condition. Following Kao et al., we expect the change to be greater in the high typicality subcondition than in the vague condition.

3. **Certainty:** We expect certainty ratings to be the lowest for the average typicality subcondition and the highest for the high typicality condition. The low typicality condition and the vague condition may or may not be signficantly different from the high typicality condition.
4. **Features inferred:** Average salience of the features inferred. We expect that number to be high (the metaphor conveys salient features of an animal.)

4 Model Evaluation

We used a handcrafted demo dataset with 10 features and 6 nonhuman animal categories (see 7.1) to run the model. When deriving salience, we fit the hyperparameters to the data, which resulted in $\mu = 0.5$ and $\kappa = 4$. We set the $n_{match} = 4$ and $n_{mismatch} = 0.25$. Following Kao et al., we set $P(c_a) = 0.01$, meaning that there is a very small but non-zero probability that the entity in question is actually not human. We also fit $\lambda = 1$ for the speaker's utility.

In all cases, regardless of the intention and typicality of a given feature, our model successfully inferred that the entity described was a person and not an animal ($P(c_p|u) = 1.00$).

Table 2. Model's highest prediction probabilities by condition

Condition	Prediction Prob.
vague	0.007 (0.002)
high typ	0.012 (0.013)
avg typ	0.004 (0.005)
low typ	0.006 (0.008)

To look at the three typicality subconditions, we consider 0 and 1 as low, 2-4 average, and 5 and 6 high typicality. The means of the model's most probable predictions are reported in Table 1. In the vague goal condition, the feature interpreted is the one associated with the highest salience (normalized typicality of 0 or 1), whereas for the specific conditions, feature interpretation depends on salience and conversational context, which in the case of average typicality push against each other.

These numbers are preliminary as they are based on a very small handcrafted dataset but the pattern suggests that the model captures the hypothesis stated in 3.3: it is easiest to interpret a metaphor if it is about a typical feature, followed by an atypical feature. Interpreting metaphor with average typicality for a feature is very difficult. Probability can be interpreted as the model's certainty of the prediction. It is the highest for high typicality, closely followed by low typicality, and noticeably lower for the vague condition and the average typicality feature-animal combinations.

Once we obtain feature priors (Experiment 3.3) and human judgments in the vague and specific goal conditions (Experiment 3.4), we will be able to calculate the correlation of the participants' certainty ratings with the model's probabilities. We also plan to run Kao et al.'s model using the obtained priors (with the priors they elicited, their model obtained a fit of r=0.6)[2]. By comparing the fit of the two models, we will be able to determine which of the two is better able to capture human understanding of metaphors. We expect that accounting for alternative utterances and relative salience of features within an animal will result in a better fit to the data.

5 Discussion

We have proposed a computational model of metaphor interpretation in the Rational Speech Act framework which combines basic principles of rational communication and communicative goals with gradient salience of features to arrive at an interpretation. We are in the process of conducting behavioral experiments to obtain feature priors and test the predictions of the model against human judgments.

The model's predictions suggest several conclusions about the nature of metaphor interpretation: most importantly, multiple factors play a role. In a vague context where the communicative goal of the speaker is not clear to the listener, the model predicts that the listener will interpret the metaphor as its most salient (which is to say, most or least typical) feature. When it is known to the listener that the speaker is trying to communicate a specific feature, the probability of that feature being interpreted increases. However, in our model, the listener is always considering both the relative salience of features within the metaphor and comparing it to the alternatives the speaker could have uttered. For example, if the listener hears "John is a fox" in response to her question "Is John scary?", the model predicts that the listener will think that the speaker is trying to communicate John's gracefulness because if she had wanted to say that John *is* or *is not* scary, she would have chosen an animal that has a high or low typicality, respectively, for scariness. However, the fact that she chose to say "fox" suggests that she wants to communicate something about John that is salient for a fox, namely gracefulness.

Klintsch et al.[1] hypothesized that a metaphor is difficult to interpret if there are very few lexical items that are strongly related to both the metaphor and the intended referent. Our model makes another prediction about what can make a metaphor hard to understand. It suggests that if the referring object has an average typicality rating (i.e., low salience) for the feature of interest, that makes the metaphor confusing and leads to misinterpretation.

We hope that the current work will make a contribution to the understanding of pragmatic metaphor interpretation. While our model captures several important features of the process at hand, there are several limitations we hope to address in future work. First, our model currently assumes that the speaker is only trying to communicate one feature but it is possible that the speaker's intention and part of the reason for choosing to use a metaphor in the first place is to communicate multiple features[2]. Secondly, currently we only consider alternative metaphorical utterances but there is, naturally, always the possibility of saying the literal version of the message which needs to be taken into account.

6 Acknowledgements

The authors would like to thank **Martin van Harmelen** for assistance with implementation and helpful insights about deriving salience from typicality and **Soren Schlassa** for lending his expertise in probability theory and assisting the authors in derivation of the L_1 equation.

7 Demo Dataset

We used the following small manually created dataset to test the model's performance and predictions. It contains typicality ratings (out of 6) for 10 features for humans and 6 non-human animal categories.

Table 3. Typicality ratings (demo dataset)

	dog	dolphin	ant	cat	ox	fox	human
strong	4	4	5	1	6	2	3
happy	6	5	3	4	2	2	3
loyal	6	4	4	2	3	1	2
sly	0	0	0	5	0	6	4
smart	4	6	4	5	2	5	4
slow	0	0	1	2	5	0	2
scary	2	0	1	2	3	3	3
lazy	1	1	0	5	0	4	5
ferocious	2	0	0	1	3	4	2
graceful	1	4	0	6	1	6	3

8 L_1 Equation Derivation

The right side of the pragmatic listener equation is derived as follows:

$$\begin{aligned}
L_1(c,f|i,u) =& \frac{L_1(i,u|c,f)\cdot L_1(c,f)}{L_1(i,u)} \\
&\propto L_1(u,f,i,c) \\
&= L_1(u|f,i,c)\cdot L_1(f,i|c)\cdot L_1(c) \\
&= S_1(u|f)\cdot L_1(f,i|c)\cdot L_1(c)
\end{aligned}$$

We assume that $L_1(u|f,i,c) = L_1(u|f,c)$ because the speaker's utterance u is only influenced by the conversational context i through f (the feature the speaker aims to communicate).

References

1. W. Klintsch and A.R. Bowles. Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, 17(4), 2002.
2. Bergen L. Goodman N. D. Kao, J. Formalizing the pragmatics of metaphor understanding. *Proceedings of the annual meeting of the Cognitive Science Society*, 36, 2014.
3. M.C. Frank and N. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084), 2012.
4. Tessler M. H. Goodman N. D. Frank M. C. Yoon, E. J. Talking with tact: Polite language as a balance between kindness and informativity. *Proceedings of the 38th annual conference of the cognitive science society*, 2016.
5. Levy R. Goodman N. D Kao, J. T. The funny thing about incongruity: A computational model of humor in puns.
6. Degen J Hawkins R. Goodman N. D. Graf, C. Animal, dog, or dalmatian? levels of abstraction in nominal referring expressions. *CogSci*, 2016.
7. M.C. Frank. Rational speech act models of pragmatic reasoning in reference games. 2016.